ABSTRACT

Title of Dissertation:	SUBSCORE REPORTING FOR DOUBLE-CODED INNOVATIVE ITEMS EMBEDDED IN MULTIPLE CONTEXTS
	Chen Li, Doctor of Philosophy, 2018
Dissertation directed by:	Associate Professor, Hong Jiao, Measurement, Statistics and Evaluation Department of Human Development and Quantitative Methodology

Reporting subscores is a prevalent practice in standardized tests to provide diagnostic information for learning and instruction. Previous research has developed various methods for reporting subscores (e.g. de la Torre & Patz, 2005; Wainer et al., 2001; Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007; Yen, 1987). However, the existing methods are not suitable for reporting subscores for a test with innovative item types, such as double-coded items and paired stimuli. This study proposes a two-parameter doubly testlet model with internal restrictions on the item difficulties (2PL-DT-MIRID) to report subscores for a test with double-coded items embedded in paired-testlets. The proposed model is based on a doubly-testlet model proposed by Jiao and Lissitz (2014) and the MIRID (Butter, De Boeck, & Verhelst, 1998). The proposed model has four major advantages in reporting subscores— (a) it reports subscores for a test with double-coded items in complex scenario structures, (b) it reports subscores designed for content clustering, which is more common than

subscores based on construct dimensionality in standardized tests, (c) it is computationally less challenging than the Multidimensional Item Response Theory (MIRT) models when estimating subscores, (d) it can be used to conduct Item Response Theory (IRT) based number-correct scoring (NCS, Yen, 1984a).

A simulation study is conducted to evaluate the model parameter recovery, subscore estimation and subscore reliability. The simulation study manipulates three factors: (a) the magnitude of testlet effect variation, (b) the correlation between testlet effects for the dual testlets and (c) the percentage of double-coded items in the test. Further, the study compares the proposed model with other underspecified models in terms of model parameter estimation and model fit.

The result of the simulation study has shown that the proposed 2PL-DT-MIRID yields more accurate model parameter and subscore estimates, in general, when the testlet effect variation is small, the dual testlets are weakly correlated and there are more double-coded items in a test. Across the study conditions, the proposed model outperforms other competing models in model parameter estimation. The reliability yielded from models ignoring dual testlets are spuriously inflated, the 2PL-DTMIRID produces higher overall score reliability and subscore reliability than models ignoring double-coded items, in most study conditions. In terms of model fit, none of the model fit indices investigated in this study (i.e. AIC, BIC and DIC) can achieve satisfactory rates of identifying the proposed true model as the best fitting model.

SUBSCORE REPORTING FOR DOUBLE-CODED INNOVATIVE ITEMS EMBEDDED IN MULTIPLE CONTEXTS

by

Chen Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Dr. Hong Jiao, Chair
Dr. Amy Hendrickson
Dr. Robert W. Lissitz
Dr. Laura Stapleton
Dr. Steven J. Ross, Dean's Representative

© Copyright by
Chen Li
2018

Dedication

To Yuchen, my beloved husband,

for the understanding, courage and love I find in you.

Acknowledgements

Completing my Ph.D. study is a challenging but rewarding experience. I feel extremely fortunate that I have received generous help and support from many people along the way. First and foremost, I would like to express my gratitude to Dr. Hong Jiao, my academic advisor and the Chair of my dissertation committee. Dr. Jiao has led my way to the field of psychometrics, coached me every step of my journey as a doctoral student and given me tremendous inspiration in research and career development. Her vision, guidance and encouragement to me will always be priceless for my future career; and her work ethic and dedication to the field of psychometrics make her a role model that I will always look up to.

Second, I would like to acknowledge my committee members. Each of you has shed lights on a different perspective of my doctoral study— Dr. Steven Ross launched my interest in language assessment into the context of quantitative research methodology and recommended me to the EDMS program; Dr. Laura Stapleton involved me in international research collaborations and provided many advice to my academic advancement; Dr. Robert Lissitz has enlightened my work at MARC and my research with his wisdom embedded in his sense of humor; and Dr. Amy Hendrickson has offered me the most inspiring internship experience that has ultimately motivated this dissertation. I sincerely appreciate every piece of advice they have given to my dissertation, to my doctoral study and to my career.

Moreover, I am very grateful for all faculty members and all my graduate colleagues in our EDMS family for being supportive. A special thanks to Dandan Liao, my dearest friend, for being my companion when dealing with similar

“doctoral-student” challenges and when tasting the same sweet bitterness of pursuing a Ph.D. degree.

I thank my parents for being my back bone throughout my entire education and especially through my Ph.D. study. I am proud to have such great parents, and I hope that my endeavors have made them very proud of me, as well.

Last but not least, I owe the most thanks to my dearest husband, Yuchen. Many people witness my achievement, but you are the only person who see my struggle, frustration and sacrifice. Thank you for gently wiping away all the negativities that I brought into our lives, having unshakable faith in me and bringing the best out of me. I thank the years of us pursuing Ph.D.s as they have bound us to the strongest team towards the challenges ahead of us.

Table of Contents

Chapter 1: Introduction

Background

In the past decade, the newly developed educational standards have put considerable emphases on acquiring higher-order cognitive skills (Krathwohl, 2002). Take the Next Generation Science Standards (NGSS) developed in 2015 as an example, they integrate three dimensions in science learning— (a) core idea, (b) crosscutting and (c) practice. The NGSS require students to demonstrate their proficiency in complex cognitive reasoning, specifically, in analyzing, evaluating and finalizing a solution to science problems through experiments. In alignment with instructional objectives, large-scale assessments also focus on evaluating higher-order cognitive skills in authentic contexts. For example, the test specification of the redesigned SAT® requires students to demonstrate their ability to "apply knowledge and skills to solve problems situated in science, social studies, and career-related contexts" (College Board, 2015).

The assessments designed to measure higher-order thinking skills often involve innovative items embedded in real-life scenarios. For example, a test item mimics real-life problem-solving processes by asking students to read a passage and listen to an audio clip before synthesizing the information from both stimuli and providing an answer. Essentially, this test requires students to synthesize information from multiple sources. Simultaneously, tasks in a test require students' knowledge from multiple sub-content domains. If a math test is intended to measure the four arithmetic operations, such a test may be composed of an item that requires students

to use knowledge in both addition and subtraction to answer the item. This item is referred to as a *double-coded item*, as it contributes to two subdomain scores.

Parshall, Spray, Kalohn, and Davey (2002) have defined innovative items as items that improve existing measures for better measurement and/or expanding measurement to new areas. Compared with a single-coded traditional item only assessing one content domain, the double-coded items nested within paired testlets are innovative for measuring content areas that require higher-order cognitive skills. Although these innovative items are advantageous in assessing proficiency and growth authentically, they impose challenges in psychometric analyses, especially in scoring students' overall performance and performance in each targeted sub-content domain. Although many types of innovative items have been developed and used in assessments, standard Item Response Theory (IRT) models are still used for psychometric analysis in practice.

An IRT model may not fit well with item response data from such double-coded innovative items with paired stimuli, due to the violation of its assumptions—local independence and unidimensionality. Local independence means that an examinee's response to one item does not relate to his/her responses to any other items in the test given his/her ability. This assumption is likely to be violated in tests consisting of testlets. A testlet refers to a bundle of items based on the same stimulus (Rosenbaum, 1988; Wainer & Kiely, 1987). When using testlets, responses to items in the same testlet are likely to be dependent due to the use of a common stimulus, given the person and item parameters. Ignoring the item clustering effect in a testlet will result in overestimated test reliability and underestimated item discrimination

parameters. (e.g. Chen & Thissen, 1997; Jiao, Wang, & Kamata, 2005; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer, Bradlow, & Wang, 2007; Wainer & Lukhele, 1997; Wainer & Wang, 2000; Yen, 1993).

The use of double-coded items, on the other hand, may violate the uni-dimensionality assumption in the application of a standard IRT model that assumes only one latent trait is assessed in the test. Compared to a multidimensional IRT (MIRT) model, fitting a unidimensional IRT (UIRT) model to response data from a multidimensional test will lead to less accurate estimates for the overall and sub-domain abilities, especially when the correlation between sub-domain abilities is high (Yao & Boughton, 2007).

Previous studies have investigated methods to accommodate testlet effects resulting from item clustering. One method is to treat dichotomous items clustered within the same stimulus as a "super item" and score the "super item" using polytomous IRT models. Instead of using the actual item response pattern, such a method estimates an examinee's ability using only the number of items answered correctly among all items in the "super item" (Wang & Wilson, 2005). Hence, treating items in a testlet as a "super item" results in loss of information and consequently undermines measurement precision. Another method models the item clustering effect directly by adding another parameter in a standard IRT model to separate the latent ability and the person-specific contextual effect for items in a testlet (e.g. Bradlow, Wainer, & Wang, 1999; Du, 1998; Jiao et al, 2005; Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005). Extending this conceptualization, a non-compensatory doubly testlet model (Jiao & Lissitz, 2014; Jiao, Lissitz & Zhan,

2017) was proposed to accommodate complex dual local item dependence (LID) for items embedded in multiple contexts such as paired passages.

Subscore reporting for a test with double-coded items is challenging. Studies have explored methods for estimating sub-domain ability using IRT models, mostly for simple structure tests where items are fully nested within latent traits (e.g. de la Torre & Patz, 2005; de la Torre & Song, 2009; Gibbons & Hedeker, 1992; Sinharay, 2010; Wainer et al., 2001; Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007; Yen,1987). Among these studies, Feinberg and Wainer (2014) conducted a study to evaluate the added-value of subscores specifically for tests with overlapping items using a MIRT model. Using a MIRT model to report subscores in the presence of double-coded items is computationally challenging, since the number of person ability parameters needing to be estimated increases drastically when the number of subdomains increases. Another challenge of using MIRT models is that since item difficulty parameters in the MIRT model cannot be decomposed for different subdomains, item parameters from MIRT models cannot be used in IRT-based number-correct scoring (NCS; Yen, 1984a). The NCS is widely used in large-scale assessments as the scores it yields are more interpretable than IRT pattern scores — the IRT-based number-correct scores are the same for students who answer the same number of items correctly in a test.

In addition, for subscores in alignment with content dimensionality instead of construct dimensionality, the use of MIRT is often hard to justify. In educational tests, subscores are often reported in alignment with either content structure of the test (more often justified by substantive or cognitive framework, thus more often

preferred by the users of the test scores) or the construct structure of the test to offer insights for different aspects of the students' general proficiency or performance. A test content dimension represents a unique content domain of a test, such as algebra or geometry in a mathematics test. Whereas a construct dimension means an unobserved/latent attribute used to describe observable behaviors measured by the test, such as listening/reading/speaking/writing ability measured in a language test (Crocker & Algina, 2008). Based on the structure for content and construct, tests can be categorized into four categories— (a) tests with unidimensional content and unidimensional construct, for example, a unit assessment on addition for a math class assess one content (i.e. addition) and only one construct (i.e. math/arithmetic operation); (b) tests with multidimensional content and unidimensional construct, for instance, the NAEP mathematics assessment is a unidimensional test (Carlson, 1993; Carlson & Jirele, 1992; Kaplan, 1995; Muthén, 1991; Rock, 1991) consisting of content sub-domains on number properties and operations, measurement, geometry, data analysis and probability, and algebra; (c) multidimensional content, multidimensional construct, a good example is a scenario-based medical licensure test that mimics the diagnosis process with "fake" standardized patients assessing various content domains and skills; and (d) tests with unidimensional content and multidimensional construct which is relatively rare in real-world assessment scenario. In an UIRT framework, all items load on one latent trait. The MIRT model, on the other hand, assumes a structure where different clusters of items load onto different latent traits. As the latent trait is unobservable, it is consistent with the concept of construct from a test development perspective. When the test is multidimensional in

content but unidimensional in construct, the use of a MIRT model for content multidimensionality may not be proper. In the case where the test is unidimensional in construct, subscores reported based on content multidimensionality will still provide beneficial diagnostic information for future instruction and learning. For such a case, the use of a unidimensional IRT model is more appropriate. Comparing to unidimensional IRT approaches, the use of MIRT is overfitting and may bring other potential problems such as lower measurement precision due to the increasing number of model parameters given the same amount of information.

Models accounting for testlet effects and estimating subdomain scores were researched and developed, respectively. No research has investigated how to estimate subscores for tests with double-coded items embedded in paired testlets (i.e., a set of items based on information from two item stems. The item stimuli can be reading passages in a reading test, or graphs and/or tables in a math or science test.). The current study is intended to develop a model to report subscores for such a test and evaluate the proposed models from various perspectives.

Purpose of the Study

This study proposes a two-parameter doubly testlet model with internal restrictions on item difficulties (2PL-DT-MIRID). The proposed model (a) accommodates complex testlet effects due to paired stimuli within multiple contexts and (b) decomposes an item difficulty parameter to difficulties that are component - specific at item level and are content domain-specific in estimating subscores. Instead of modeling content domain scores as multidimensional, as in MIRT models for examinees' domain ability estimation, the proposed model reports scores for each

content domain by decomposing item parameters, especially the item difficulty parameters, into domain-specific item difficulties for subscore estimation within a UIRT modeling framework. This formulation is consistent with the situation where subscores represent content multidimensionality. The item parameters of the proposed model can also be used to estimate examinees' domain abilities using IRT-based NCS (Yen, 1984a).

A Monte Carlo simulation study is conducted to evaluate the proposed model in modeling item responses from an arithmetic test. In this simulation study, dichotomous item responses are generated based on the proposed 2PL-DT-MIRID to mimic an arithmetic test with double-coded items embedded in multiple contexts. The simulated item responses to double-coded items embedded in paired testlets are scored by (a) the proposed 2PL-DT-MIRID, (b) the testlet MIRID (c) the MIRID (Butter, De Boeck, & Verhelst, 1998), (d) a two-parameter doubly testlet model (2PL-DTM; Jiao & Lissitz, 2014; Jiao et al. 2017), (e) the unidimensional two-parameter logistic (2PL) IRT model, and (f) the IRT-based NCS using model parameters from the proposed model. The performance of the proposed model is evaluated in comparison with the other scoring methods in terms of parameter estimation accuracy and reliability of the estimated subscores and the total scores.

This study addresses the following three research questions.

1. How well can the parameters of the proposed 2PL-DT-MIRID be recovered across different study conditions? In other words, how do the manipulated factors (i.e. the magnitude of the testlet effect represented by the standard deviation of testlet effects, the correlation between the testlet

effects of the two testlets and the percentage of double-coded items in the test) impact the model parameter recovery for the proposed model?

2. How do ignoring the dual testlet effects, ignoring the testlet effect and/or ignoring the effect of the double-coded items impact the model parameters recovery, the subscores estimation accuracy, and the overall score and subscore reliability?
3. Which model fit index is more capable of identifying the true model across different study conditions?

Significance of the Study

Since the adoption of the No Child Left Behind (NCLB) Act in 2001, the reporting of subdomain test scores for diagnosis to address specific instructional goals has been encouraged. As such, reporting subscores in addition to the summative test scores has become increasingly prevalent in educational assessment. This advocacy is rooted in the main advantage of subscores—their capability of providing diagnostic information for learning and instruction (Sinharay, Puhan, & Haberman, 2011). For example, a summative mathematic score is usually reported along with domain scores, such as number sense, algebra, geometry and data analysis, to improve learning and instruction.

The emphasis of academic content knowledge and non-cognitive skills in educational standards requires task-based and context-based items to make assessment authentic. Although research has indicated that double-coded items cannot help to increase the added-value of subscores (e.g., Feinberg & Wainer, 2014), the use of such items is inevitable, as they are effective in assessing complex cognitive

skills and intertwined knowledge network embedded in real world problem-solving. For example, SAT® reports two section scores, two cross-test scores, three test scores and 7 subscores, in addition to the test total score. Among the subscores, some scores are obtained based on double-coded items.

The current practice in subscore reporting is that a double-coded item is treated as a single-coded item and counted twice for two subscore computation, one for each subscore. The estimation of one sub-content domain score is contaminated with information from the other sub-content domain. The proposed 2PL-DT-MIRID decomposes item difficulty for the double-coded items into domain-specific ones, hence, the estimation of domain ability is purified. This model is formulated to accommodate tests targeting multidimensional content areas but unidimensional constructs where the use of MIRT is hard to justify. In addition, the model parameters calibrated from the proposed model can be used not only in pattern scoring, but also in IRT-based NCS. It is more flexible than MIRT models where IRT-based NC domain scores cannot be estimated using the item parameter estimates from a MIRT model.

Overview of the Dissertation

This dissertation contains five chapters. The first chapter describes the problem to be investigated in this dissertation research. In Chapter 2, previous studies on testlet IRT models, psychometric models used for subscore reporting, and the MIRID are reviewed and summarized to scaffold for the proposal of the non-compensatory 2PL-DT-MIRID. Specifically, reviewing studies on item clustering effects and subscore reporting (a) presents the development of available methods of

reporting subscores for tests containing testlets, and (b) justifies the necessity of the proposed model based on the limitation of the available ones. Further, the synthesis of studies on testlet models and the MIRID lays a theoretical foundation for the formulation of the proposed model. Chapter 3 first presents the formulation of the proposed model, then outlines the simulation study investigating model parameter recovery and model selection issues. Information related to simulation conditions (i.e. fixed factors and manipulated factors), data generation, model identification and estimation, methods for summarizing the results, and model evaluation criteria is presented. Simulation conditions are justified by both theoretical evidence and results from pilot studies. Chapter 4 summarizes the study results in terms of parameter estimation, score reliability and model selection. Key findings are highlighted with tables and figures in this section. Results of the simulation study are further discussed and synthesized in comparison and contrast with previous investigations on relevant topics in the last chapter. A summary of contribution and limitation of this study concludes Chapter 5 and this dissertation.

Chapter 2: Theoretical Framework

This chapter reviews and synthesizes previous studies to present the inspiration, motivation, and theoretical framework for the proposed model. Literature reviewed is categorized into three focused topics in this chapter – (a) IRT models and testlet response theory (TRT) models for local item dependence, (b) methods and models developed for subscore reporting, and (c) the MIRID. The first section of this literature review introduces IRT models, their assumptions and the development of TRT models based on the IRT models. In the second section, the methods and models used for subscore reporting are discussed, including the advantages and disadvantages of the available methods. As the MIRID is an important component in the proposed model, the model formulation, parameter estimation and the application of the MIRID are introduced in detail in the third section.

Item Response Theory and Testlet Response Theory Models

Item Response Theory Models

Classical Test Theory (CTT) and Item Response Theory (IRT) are two widely used measurement theories in measuring test performance (in CTT) or latent trait (in IRT) and instrument development. The CTT defines an examinee's observed score as the sum of his/her true score and error score. The error score is attributable to random and systematic mechanisms. The CTT is carried out with weak assumptions— (a) an examinee's error score does not correlate to the his/her true score, (b) the error scores for an examinee on parallel forms are not correlated and (c) the average error score in a population of examinees is 0 (Hambleton & Jones, 1993). In CTT, an examinee's

true score is defined as the expected value of the observed scores across parallel forms where tests measure the same construct with equal size of measurement error.

IRT is a family of statistical models with logit or probit link functions, which characterizes the relationship between an examinee's performance on an item and his/her latent ability on the targeted construct/content measured by items in the test (Hambleton & Jones, 1993). In an IRT model, the probability for an examinee obtaining a correct answer to an item is modeled as a mathematical function of the examinee's latent ability and item characteristics, such as item difficulty and item discrimination. Given that the selected IRT model reflects the true relation between the item responses and the latent ability, item statistics and person abilities yielded from the IRT model are invariant by directly modeling the relative standing of examinee's ability and item difficulty at the same time. Invariant parameter estimation of an IRT model is a big advantage compared with CTT. In CTT, the estimates of the true scores are entirely dependent on the test, that is, the estimates are inconsistent across tests; and item statistics (e.g., item difficulty and item discrimination) are sample dependent.

A variety of IRT models have been developed for fitting different types of item responses to different types of items. The choice of an appropriate IRT model depends on (a) whether item responses are dichotomous or polytomous, (b) if the response categories are ordered, (c) how many abilities contribute to the performance on an item, and (d) the relationship between the item responses and the underlying ability(ies) (Hambleton & Jones, 1993). Since the functional form of the proposed 2PL-DT-MIRID is based on a unidimensional IRT model for dichotomously scored

item responses, commonly used unidimensional dichotomous IRT models are briefly introduced as follows.

The one-parameter logistic (1PL) IRT model was developed based on the framework of the generalized linear regression model with parameters interpreted in the context of measurement (e.g., Andrich, 2004; Linacre, 2005). Mathematically, as presented in Equation 1, the probability of examinee j obtaining a correct answer to item i, denoted as $P_{ij}(X_{ij} = 1)$, is a logistic function of the difference between the examinee j's latent ability, denoted as θ_j, item difficulty, denoted as b_i, and item discrimination, denoted as a. Item difficulty is the point on the ability scale where the probability of getting the item correct is 50% for IRT models with no upper or lower asymptotes. The 1PL IRT model assumes that all items differ only in terms of item difficulty while equally discriminating for all examinees.

$$P_{ij}\left(X_{ij} = 1 \middle| \theta_j, a, b_i\right) = \frac{1}{1 + \exp\left[-a\left(\theta_j - b_i\right)\right]}. \quad (1)$$

The Rasch model (Rasch, 1960) is also frequently used in test operational practice. It is very similar to the 1PL IRT model, except that the Rasch model constrains item discrimination to be 1 for all items in the test.

A two-parameter logistic (2PL) IRT model was proposed by Birnbaum (1968) to allow items to differ on item discrimination in addition to difficulty. The 2PL IRT model proposed is presented in Equation 2, where the item discrimination parameter is denoted as a_i.

$$P_{ij}\left(X_{ij} = 1 \middle| \theta_j, a_i, b_i\right) = \frac{1}{1 + \exp\left[-a_i\left(\theta_j - b_i\right)\right]}, \quad (2)$$

Conceptually, an item with higher discrimination power will have a steeper slope on the Item Characteristic Curve (ICC), an S-shaped curve depicting the

probabilities of obtaining a correct answer along the latent ability scale. When item discrimination is higher, examinees with abilities closer to item difficulty are better separated into different ability levels (Hambleton, Swaminathan, & Rogers, 1991). In other words, the higher the discrimination power, the more informative the item. The item discrimination is denoted as a_i that is defined by the slope of ICC at the inflection point. Although the theoretical range of the item discrimination parameter is from negative infinity to positive infinity, negative item discrimination is considered as a red flag for a "bad" item. The flagged items are usually discarded after analyzing item responses from field tests. This is because, when item discrimination is negative, examinees with higher abilities will have lower probabilities of getting an item correct. A common range of item discrimination in testing practice is (0, 2) (Hambleton et al, 1991).

For computational simplicity, Birnbaum (1968) used a scaling parameter D ($D \approx 1.7$) to minimize the difference between the logistic ogive and the normal ogive (Camilli, 1994). See Equation 3.

$$P_{ij}\left(X_{ij} = 1 \middle| \theta_j, a_i, b_i\right) = \frac{1}{1 + \exp\left[-Da_i\left(\theta_j - b_i\right)\right]}. \quad (3)$$

The ICCs for the Rasch model, the 1PL and the 2PL IRT models have a lower asymptote of 0, which means that the probability of answering the item correctly will asymptotically approach to 0 as the person ability approaches negative infinity. Thus, these models ignore the possibility of obtaining a correct answer only by chance or other plausible factors, such as pre-knowledge. Birnbaum's (1968) three-parameter logistic (3PL) IRT model incorporates a pseudo-guessing parameter to increase the

lower asymptote of the ICC allowing random guessing effect. The 3PL IRT model is presented as in Equation 4.

$$P_{ij}\left(X_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+\frac{1-c_i}{1+\exp\left[-a_i\left(\theta_j-b_i\right)\right]}, \quad (4)$$

where c_i represents the lower asymptote for the ith item.

There are three underlying assumptions for an IRT model. These assumptions guarantee accurate parameter estimation and the valid inferences obtained from an IRT model by addressing (a) the model data fit, (b) dimensionality and (c) local independence.

Model data fit. The ICC specified based on the functional form of the IRT model should reflect the true relation between the observed item response and the latent person ability. First, the specified IRT model (a) should have a functional form that is monotonically increasing to describe the intrinsic relationship between latent ability and the probability of obtaining a correct answer— the higher the ability, the higher the probability of getting an item correct. Second, the model should include an appropriate number of item parameters that present the relations between item responses and the item characteristics, for example, a 3PL IRT model may be used for multiple-choice items where it is possible to get a correct answer by random guessing.

Dimensionality. The person ability parameter(s) in an IRT model demonstrate(s) the dimension(s) on which examines are measured by items in the tests. The UIRT model, as its name suggested, models the situation where an item only measures one underlying ability. In other words, only one latent trait underlies the item response (Hambleton et al., 1991). Yet, research has shown that the uni-dimensionality assumption is often challenged in real-world assessment scenarios

(Ackerman, 1994; Nandakumar, 1994; Reckase, 1985). In some cases, the number of dimensions of the ability is underestimated (Reckase & Hirsh, 1991).

Test multidimensionality could be either planned and/or unintended. The most common motivation for designing a multidimensional test is content clustering. For example, a test on science may test students' knowledge in many sub-content domains, such as physics, chemistry and biology, etc. In other cases, multidimensional trait may also entail multidimensionality of a test. One example is a language test that assesses four skills of language proficiency— listening, speaking, reading and writing. On the other hand, unintended multidimensionality can be induced by test speediness (Lu & Sireci, 2007), passage dependency (e.g. Bradlow et al., 1999; Hartig & Höhler, 2009), and/or item format (Yao & Schwarz, 2006).

A planned multidimensional test can have a simple structure or a complex structure. A simple structure refers to the situation where each of the items in the test only contributes to one latent trait, and the test assesses more than one latent trait (e.g. Lee & Brossman, 2012; Reckase, 2009). In other words, a simple structure test is a multidimensional test consisting of two or more unidimensional sub-tests (Wang et al., 2006). As items are fully nested within the ability dimensions, such structure is also referred to as between-item multidimensionality (e.g., Adams, Wilson & Wang, 1997; Hartig & Höhler, 2008; te Marvelde, Glas, Van Landeghem, & Van Damme, 2006). A complex structure, also referred to as within-item structure, allows items in a multidimensional test to load on more than one latent trait (e.g. Lee & Brossman, 2012; Reckase, 2009). Diagram representations of a simple structure and a complex structure are presented in Figure 1. In Figure 1, items 3 and 4 in the complex structure

test load on both latent traits in the model, whereas all items in the simple structure measure only one latent trait.

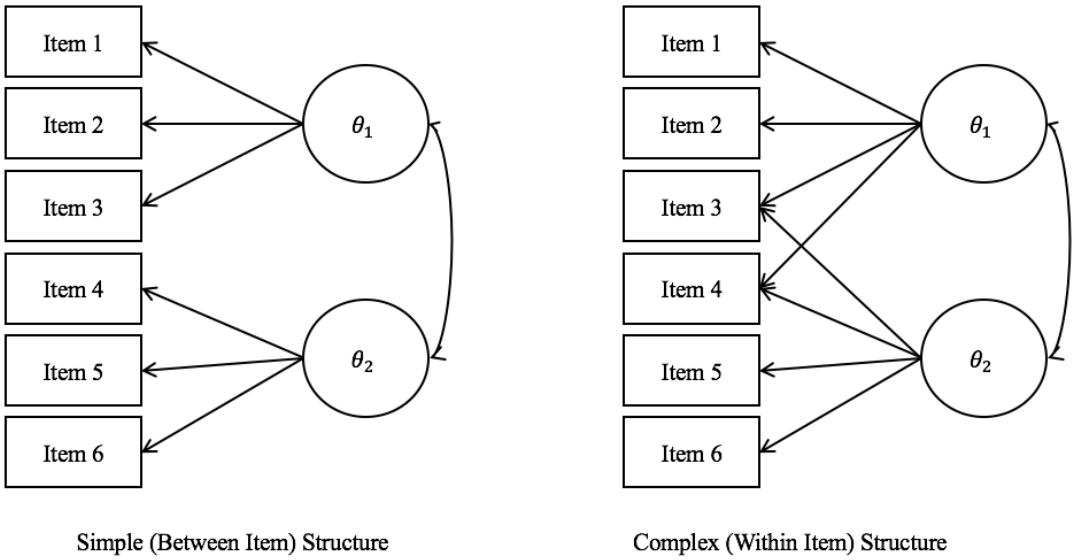


Figure 1. Examples of a simple structure model (left) and a complex structure model (right)

Multidimensional IRT (MIRT) models have been developed to model item responses based on items assessing more than one dimension of abilities (e.g., Reckase, 1997, 2009). The application of the MIRT models will be discussed together with other methods used for subscore reporting later in this chapter. As the current study focuses on reporting subscores by decomposing item difficulties, the uni-dimensionality of the latent trait is assumed in the proposed model.

Local independence. The assumption of local independence in an IRT model consists of two facets—(a) local person independence (LPI) and (b) local item independence (LII, Reckase, 2009, p. 13). LPI is achieved when item responses to a specific item are uncorrelated after controlling for the persons' abilities. Mathematically, the LPI can be expressed as

$$P(X = \boldsymbol{x}|\boldsymbol{\theta}) = \prod_{j=1}^{n} p(x_{ij}|\theta_j)\,, \tag{5}$$

where the probability of a group of n examinees with ability vector $\boldsymbol{\theta}$ obtaining a specific response pattern $\boldsymbol{x}$ on item i, is the product of the probability of each person with an ability parameter of θ_j getting a response of x_{ij} on item i. This assumption is often violated due to person clustering due to factors such as cluster sampling, group intervention and problem-solving strategies. (e.g. Jiao, Kamata, Wang, & Jin, 2012). As the current study does not involve LPI, this paper will not discuss LPI in further details.

The LII describes the situation where an examinee's response to one item does not relate to his/her response to another item after controlling for his/her latent ability. LII is described mathematically as in Equation 5— the probability of getting a specific response pattern on a test with I items for person j with ability θ_j is the product of the probability of answering each item in the test correctly.

$$P\big(X = \boldsymbol{x}\big|\theta_j\big) = \prod_{i=1}^{I} p\big(x_{ij}\big|\theta_j\big). \tag{6}$$

When evaluating local independence of an IRT model, it is necessary that both LII and LPI are satisfied.

Yen (1993) identified 10 causes for LID— (a) external assistance/interference, (b) speededness, (c) fatigue, (d) practice, (e) item or response format, (f) passage dependence, (g) item chaining, (h) explanation of previous answer, (i) scoring rubrics or rater clustering and (j) content, knowledge and abilities clustering. This study focuses on modeling LID caused by passage dependence, that is, item clustering

effect introduced by the use of common stimuli, such as reading passage, graph and/or tables.

Consequences of ignoring LID. Previous research has identified three major problems of ignoring LID in the use of an IRT model—(a) biased item and person parameter estimation, (b) inflated reliability estimation, and (c) equating errors.

First, biased estimates of item and person parameters are induced by LID. Chen and Thissen (1997) found biased item discrimination parameter estimates with the presence of LID, but their study did not conclude on the direction of the biased item parameters. Ackerman (1987) and Reese (1995) both found that the item discrimination parameters were over-estimated in the presence of LID. On the other hand, Bradlow et al. (1999) and Wainer et al. (2000) discovered that the item discrimination parameters were underestimated when ignoring the item clustering effect caused by testlets. They also showed that the discrepancy between the item discrimination estimates yielded from the underspecified models and those from the true model was larger when LID is larger. The seemingly contradictory conclusion on the direction of bias in item discrimination estimates is attributable to the difference in formulating LID in data generation. Item discrimination indicates the correlation between item response and person ability. In studies by Bradlow et al. (1999) and Wainer et al. (2000), LID is modeled by adding a person specific testlet effect to a standard IRT model, the true item discrimination is a measure of the correlation between the item responses and the combination of ability and testlet effect. Ignoring testlet effect naively assumes the variance in item responses is due to person ability solely. Consequently, the "true" response-ability correlation is undermined, and so is

item discrimination (Bradlow et. al., 1999). Whereas in Ackerman (1987) and Reese (1995), LID is simulated directly by correlating item responses, which lowers the total noise contained in the item responses. Consequently, item discrimination is overestimated in such studies. When ignoring LID, item difficulty estimates were found attenuated towards the mean (Ackerman, 1987; Bradlow et al., 1999; Reese, 1995) and the pseudo-guessing parameters are underestimated (Reese, 1995). Yet, the bias in item difficulty and the pseudo-guessing parameter estimates are not as "alarming" (Reese, 1995, p. 10) as those in item discrimination parameter estimates. Further, Ackerman (1987) and Reese (1995) also indicated that ability estimates tend to be biased at the lower and upper ends of the latent ability scale as LID increases. Specifically, the lower abilities are underestimated, and the higher abilities are overestimated when ignoring LID.

In the IRT framework, item and test information are used as an index for measurement precision of an item and a test, respectively. Information, standard error of measurement (SEM) and reliability are conditional on the latent ability level in IRT. The information of an item i, at ability level θ, is defined in Equation 7 where the probability of getting the item right is denoted as $P_i(\theta)$, the probability of obtaining an incorrect answer, denoted as $Q_i(\theta)$, and $P_i'(\theta)$ is the derivative of $P_i(\theta)$.

$$I_i(\theta) = \frac{(P_i'(\theta))^2}{P_i(\theta)Q_i(\theta)}. \quad (7)$$

For example, if the relation between the latent ability, denoted as θ, and the item response to item i is modeled by Rasch model, the item information is the product of the probability of answering the item correctly, denoted as $P_i(\theta)$, and that of providing an incorrect answer, denoted as $Q_i(\theta)$. That is,

$$I_i(\theta) = P_i(\theta)Q_i(\theta). \quad (8)$$

The test information at a latent ability level is the sum of all item information at that latent ability. The SEM at a specific ability level θ is the inverse of the square root of the item information at ability level θ. Conceptually, higher information leads to lower SEM and higher precision of the measurement. Reliability indicates measurement precision. In other words, reliability represents to what extent the measurement is without error. Therefore, as the SEM increases the reliability decreases. As the information can be conceptualized as an index evaluating how well the item distinguishes students with high and low abilities, the changes in information and reliability are in the same direction.

Studies have shown evidence that item and test information are inflated when LID is present (Ip, 2000; Thissen, Steinberg, & Mooney, 1989; Reese, 1995; Wainer & Wang, 2001). The SEM is also found to be underestimated when LID is present but ignored (Wainer, 1995; Wainer & Thissen, 1996). Consistently, studies have also observed that reliability is overestimated when ignoring LID (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Lukhele, 1997; Wainer & Thissen, 1996, Yen, 1993; Zenisky, Hambleton, & Sireci, 2002).

In addition, Yen (1984b) found substantial unsystematic errors in equating test forms when LID is present. As equating is not the focus of the current study, the impact of LID on test equating will not be discussed in detail.

Testlet Response Theory Models

Rosenbaum (1988, p. 349) proposed the concept of *item bundle* as "a small group of multiple choice items that share a common reading passage or graph, or a

small group of matching items that share distractors". Wainer and Kiely (1987) discussed the use of a *testlet* in the context of Computer Adaptive Testing (CAT), where items related to the same content area are developed as a unit and arranged hierarchically or linearly with predetermined paths for students with different abilities. Although these concepts were discussed for different test scenarios, they were both proposed for modeling the effect of item clustering—the violation of LII assumption (Rosenbaum, 1988; Wainer & Kiely, 1987) and accommodating the contextual effect (Wainer & Kiely, 1987). In general, the impact of LID due to testlets is not negligible (e.g., Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999; Lee, 2004; Sireci et al., 1991; Thissen et al., 1989). In this study, a testlet refers to a cluster of items that are nested within the same item stimuli. Due to the common stimuli, a student's responses to all items in the cluster are more likely to be impacted by the same content or context, a construct or content irrelevant factor, which may introduce noise in estimating the latent ability.

As described in Chapter 1, the authentic and scenario-based assessments have led to the popularity of testlets in large-scale standardized assessment. For example, reading ability is often assessed by asking students to answer a group of multiple-choice items based on a passage. The same format is also common in a math test where students are required to solve a series of computational questions with information given in a table or a graph to demonstrate their mastery of relevant math concepts. In the presence of a testlet, using a basic IRT model is problematic. Some psychometric models were developed to address the insufficiency of basic IRT models in modeling item responses yielded from a test with testlets.

Bradlow et al. (1999) proposed a Bayesian random effects model for testlets by adding a random effects parameter to the standard 2PL IRT model to account for dependence among items that are nested within the same testlet. The baseline model they used in their study is a probit version of the logit 2PL IPT model. The mathematical formulation is presented as in Equation 9.

$$t_{ij} = a_i(\theta_j - b_i - \gamma_{jd(i)})) + \epsilon_{ij},$$
where $$y = \begin{cases} 1 \ , \ t_{ij} > 0 \\ 0 \ , \ t_{ij} \leq 0 \end{cases}. \quad (9)$$

In Equation 9, $\gamma_{jd(i)}$ represents the person specific testlet effect for person j on item i nested in testlet d. It is assumed that $\gamma_{jd(i)}$ follows a normal distribution with a mean of 0 and a variance of σ_γ^2. The sum of $\gamma_{jd(i)}$ across all people equals to 0. The dependency of the items is accommodated in this formulation as the testlet effect for person j is assumed to be the same for all items in the same testlet. Such a testlet response model can also be formulated as a logistic model. In Equation 10, the probability of answering an item correctly is a function of person j's ability, denoted as θ_j; item discrimination a_i, item difficulty b_i and the testlet effect, denoted as $\gamma_{jd(i)}$. In this model, item discrimination for each item is considered to be the same for the general ability and the testlet effect.

$$P_{ij}\left(X_{ij} = 1 \middle| \theta_j, a_i, b_i\right) = \frac{1}{1 + \exp\left[-a_i\left(\theta_j - b_i - \gamma_{jd(i)}\right)\right]}. \quad (10)$$

The model proposed in Bradlow et al. (1999) has been extended to a three-parameter logistic testlet model (Du, 1998; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002) by adding the pseudo-guessing parameter. The 3 PL Testlet Response Theory (TRT) model is presented in Equation 11.

$$P_{ij}\left(X_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+\frac{1-c_i}{1+\exp\left[-a\left(\theta_j-b_i-\gamma_{jd(i)}\right)\right]}, \quad (11)$$

where c_i represents the probability of obtaining a correct response to item i by chance. Other parameters are interpreted the same as in the 2PL TRT model.

Wang and Wilson (2005) proposed the Rasch model based TRT model for the situation where examinees have no random chance of obtaining the right answer for an item (i.e. $c_i = 0$) and the items are equally discriminant with item discrimination being 1 (i.e. $a_i = 1$). Wang and Wilson (2005) consider the Rasch TRT model as a special case of the multidimensional random coefficients multinomial logit model. To model the testlet effect, this model incorporated new dimensions for testlets in addition to the general ability dimension. In other words, all items load on the general factor (i.e. general ability), items clustered within a testlet load on an additional factor specifically for the testlet to account for the testlet effect. The Rasch TRT model is presented in Equation 12.

$$P_{ij}\left(X_{ij}=1\middle|\theta_j,b_i\right)=\frac{1}{1+\exp\left[-\left(\theta_j-b_i-\gamma_{jd(i)}\right)\right]}. \quad (12)$$

From the perspective of multilevel modeling, Jiao, Wang and Kamata (2005) proposed a three-level one-parameter logistic testlet model in which item effects are modeled at level-1, item clustering effects are modeled at level-2, and person effects are modeled at level-3. Such a model is presented in Equation 13.

$$\eta_{itj}=\log\left(\frac{p_{itj}}{1-p_{itj}}\right).$$

At level 1: $$\eta_{itj}=\beta_{0tj}+\beta_{1tj}X_{1itj}+\cdots+\beta_{(k-1)tj}X_{(k-1)itj},$$

At level 2: $$\begin{cases}\beta_{0tj}=\gamma_{00j}+u_{0tj}\\ \beta_{1tj}=\gamma_{1t0}\\ \cdots\\ \beta_{(k-1)tj}=\gamma_{(k-1)t0}\end{cases}, \quad (13)$$

At level 3:
$$\begin{cases} \gamma_{00j} = \pi_{000} + r_{00j} \\ \gamma_{10j} = \pi_{100} \\ \quad \dots \\ \gamma_{(k-1)0j} = \pi_{(k-1)00} \end{cases},$$

where p_{itj} is the probability that person j responds to item i in testlet t correctly, with $i = 1, \dots, k-1$ (the k^{th} item is the reference item), $j = 1, \dots, J$ and $t = 1, .. T$; η_{itj} is the logit of p_{itj}; and at Level 1, X_{qitj} $(q = 1, \dots, k-1)$ represents the q^{th} dummy coded variable for person j , with value 1 when $q = i$ and 0 when $q \neq i$. For item i in testlet t; the coefficient β_{0tj} is the intercept which represents the effect of the reference item, and β_{qtj} is the coefficient associated with X_{qitj}, which represents the effect of the q^{th} item relative to the reference item; at Level 2, u_{0tj} is a random component of β_{0tj} and is distributed as $N\ (r_{00j}, \sigma_u^2)$, indicating the person-item cluster interaction for person j and testlet t, the random effect u_{0tj}, is analogous to $\gamma_{id(i)}$ in Bradlow et al. (1999); at level-3, γ_{00j} is decomposed to the average person ability and the person-specific effect r_{00j}; the person effect follows a normal distribution with a mean of 0 and standard deviation of σ_r^2.

Alternative methods applied the conceptualization of a bi-factor model in the context of a testlet IRT model allowing the item discrimination parameters to be different for testlet and general ability dimensions to allow more flexibility and generality (Li, Bolt, & Fu, 2006; Tao, Xu, Shi, & Jiao, 2013). For example, Li et al. (2006) proposed a more generalized model (see in Equation 14),

$$p_{ij}\left(X_{ij} = 1 \middle| \theta_j, a_{i1}, a_{i2}, b, \gamma_{jd(i)}\right) = \frac{1}{1 + \exp\left[-\left(a_{i1}\theta_j - b_i + a_{i2}\gamma_{jd(i)}\right)\right]}, \quad (14)$$

where a_{i1} and a_{i2} are item discrimination parameters for the general ability θ_j and testlet effect $\gamma_{jd(i)}$, respectively. Allowing item discrimination to differ for ability and testlet-ability interaction provides a more flexible modeling approach in modeling the relation between item responses and items nested within testlets. This model assumes that both person ability θ_j and testlet effect $\gamma_{jd(i)}$ follow a standard normal distribution and are independent of each other.

Recently, a non-compensatory doubly testlet model (Jiao & Lissitz, 2014; Jiao et al., 2017) was proposed to accommodate complex dual LID for items embedded in multiple contexts such as paired passages based on a state reading test. The 2PL-DTM (Jiao & Lissitz, 2014; Jiao et al., 2017) is presented as follows.

$$P(u_{ij}=1) = \left(\frac{1}{1+\exp\left(-a_i\left(\theta_j - b_i + \gamma_{jd_1(i)}\right)\right)}\right) * \left(\frac{1}{1+\exp\left(-a_i\left(\theta_j - b_i + \gamma_{jd_2(i)}\right)\right)}\right), \quad (15)$$

where the probability of obtaining a correct answer for an item embedded in paired contexts correct is a function of the examinee's ability, denoted as θ_j, item difficulty, b_i, item discrimination, a_i , and testlet effects from each stimulus, denoted as $\gamma_{jd_1(i)}$ and $\gamma_{jd_2(i)}$, respectively. The non-compensatory relation indicates the necessity of mastering both testlets content to answer an item embedded in multiple contexts correctly. However, their model does not estimate subdomain abilities in tests with double-coded items in paired passages. The proposed model is based on the 2PL IRT model and the non-compensatory doubly testlet model (Jiao & Lissitz, 2014) described in this section.

Subscore Reporting

Test scoring is to provide a numeric summary of an examinee's performance on the test by summarizing his/her response to each individual item in the test (Thissen & Wainer, 2001). In modern measurement theory, an examinee's responses to items in a test are considered as indicators of his/her underlying trait or traits. In addition to a summative score, subscores are also reported in many tests. A subscore is also referred to as a domain score, an objective score, a skill score, subscale score or a diagnostic score. It indicates examinees' levels of mastery or proficiency in a sub-category of the holistic trait assessed in the test based on a subset of items in the test. A subscore serves two major functions in the learning-assessment dynamics— first, it indicates an examinee's strength and weaknesses; second, it helps examinees to work harder on the area(s) that he/she performed poorly and to make progress in future learning (e.g., Boughton, Yao, & Lewis, 2006; Thissen & Edwards, 2005; Yen, 1987).

Promoted by educational policy such as NCLB, reporting subscores to assist instruction and learning is increasingly prevalent in various testing programs. For example, a standard SAT® test (i.e. a SAT® test excluding the optional Essay test) reports 15 scale scores to each examinee, including 1 total score, 2 section scores, 2 cross section scores, 3 test scores and 7 subscores. Another example is a state writing portfolio test that reports subscores on planning, drafting, revising, editing, structure, ideas and language use. This section reviews literature on psychometric approaches for reporting subscores— from CTT based methods to UIRT methods, then to MIRT approaches. This section will briefly discuss each of the methods first, then highlight

the advantages and disadvantages of these methods by summarizing comparison studies of these methods. The reviewed studies have used a variety of terminologies to refer to the concept of a subscore, including subscale ability, objective score/ability, subdomain ability, ability dimension, dimensional ability and dimensional latent trait. In this section, the word "subscore" will be used consistently to avoid confusion on terminology.

Classical Test Theory Based Approaches for Subscore Reporting

Raw subscore. A raw subscore is also called a CTT-based number-correct subscore. As its name suggests, a raw subscore is obtained by summing up the coded outcomes across a subset of items that are designed to assess students' proficiency in a sub-content domain based on test specification. For example, an 8th grade state math test with nationally-aligned standards has 4 sub-content areas—algebra, geometry, data and number/computation (Embretson & Yang, 2013). In this test, each item assesses only one sub-content area. If raw subscores are computed from such a test, four raw subscores would be reported. Each subscore is the sum of the item scores across all items assessing the same sub-domain area.

Kelley's univariate regression. Kelley's univariate regression method (Kelley, 1947) is a regression-based method for estimating scores. It uses the group mean, the observed score and the reliability of the score to predict the true score. Kelley's equation is presented in Equation 16.

$$\widehat{\tau} = \rho X + (1 - \rho)\mu, \tag{16}$$

where $\hat{\tau}$, the predicted true score, is regressed on the observed score, denoted as X and the mean score of the group, weighted by ρ, the score reliability, and $1 - \rho$,

respectively. Conceptually, this method estimates the true score as a composite of the reliable part of the observed score and a complement proportion of the group mean to remove the unreliable part of the observed score. As the reliability of the test increases, the predicted true score is pulled towards the observed score; if test scores are not reliable at all (i.e. $\rho = 0$), the predicted true score is the average score across the group of examinees. This method assumes that each item contributes to only one content area. Kelley's equation can be easily applied to subscore reporting for a test (Skorupski & Carvajal, 2010).

Haberman (2008) proposed three variations of Kelley's original method to predict a true subscore using the observed subscore, observed total score and a weighted average of the observed subscores and the observed total score. Research found that, the weighted average of the observed subscore and the total score is a better predictor of the true subscore (Feinberg, 2012; Haberman, 2008; Puhan, Sinharay, Haberman, & Larkin, 2010).

Yen's Objective Performance Index (OPI). Yen's OPI (1987) is proposed to stabilize the estimates of the subscores. This method estimates an examinee's performance on an objective area with information of his/her overall performance on a test using Bayesian estimation with a beta prior incorporating IRT item and person parameters to obtain OPI as the mean of the posterior distribution. In Yen (1987), a subscore is referred to as an objective score. Yen's OPI (1987) is presented as follows.

$$\tilde{T} = w_j \widehat{T}_J + \left(1 - w_j\right) \frac{x_j}{n_j}, \quad \text{where} \quad w_j = \frac{n_j^*}{n_j^* + n_j} \text{ and } \widehat{T}_J = \frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}(\hat{\theta}). \tag{17}$$

Like Kelley's method, the OPI can be expressed as a weighted sum of the estimated proportion-correct score, denoted as $\hat{T}_j$, across I items for objective j given estimated overall ability $\widehat{\theta_j}$, and the observed proportion-correct score, that is, the quotient of observed score for objective j, denoted as x_j, and the number of items assessing objective j, denoted as n_j. The weight w_j is calculated as the ratio of theoretical number of items in the objective, denoted as n_j^* and the sum of the theoretical number of items and the observed number of items in the objective. The theoretical number of items is derived based on the distribution of prior information (e.g., grade in school or test scores from another test) and item response function.

Wainer et al.'s augmented subscore. Wainer et al. (2001) proposed an approach to estimate a CTT-based subscore with information from other subscores. This method is a multivariate version of Kelley's univariate regression method (Kelley, 1947). The mathematical formulation of Wainer et al.'s augmented subscore is presented in Equation 18.

$$\hat{\boldsymbol{\tau}} = \overline{\boldsymbol{x}} + \boldsymbol{B}(\boldsymbol{x} - \overline{\boldsymbol{x}}), \quad \text{where} \quad \boldsymbol{B} = \Sigma_{true}\Sigma_{obs}^{-1}. \tag{18}$$

In Equation 18, $\hat{\boldsymbol{\tau}}$ is a vector of the predicted subscores for a test, $\overline{\boldsymbol{x}}$ is a vector of means for all subsets of items, $\boldsymbol{x}$ is a vector of observed scores. $\boldsymbol{B}$ is a matrix of reliability-based regression coefficients which is obtained by multiplying covariance matrix of the true scores and the inverse of the covariance matrix of the observed scores. The population covariance matrix of the true scores and the observed scores, denoted as Σ_{true} and Σ_{obs} can be estimated as sample covariance matrices, $\boldsymbol{S}_{true}$ and $\boldsymbol{S}_{obs}$. $S_{true}^{vv'}$ is the vv' element in $\boldsymbol{S}_{true}$, and $S_{obs}^{vv'}$ is the vv' element in $\boldsymbol{S}_{obs}$. When $v \neq v'$, $S_{true}^{vv'} = S_{obs}^{vv'}$. Since CTT assumes that error scores are not correlated with true

scores, the covariance of two true scores should be equal to the covariance of the two observed scores. When $v = v'$, $S_{true}^{vv'} = r_v \, S_{obs}^{vv'}$, where r_v is the reliability for the v^{th} subset of items. Cronbach's alpha as the lower bound of reliability is normally used in the computation, that is, $r_v = \alpha$.

Kelley's regression method (1947), Yen's OPI (1987) and Wainer et al.'s (2001) augmented subscore method are all augmented subscores as they all used auxiliary information to estimate subscores in addition to an examinee's performance on the subset of items for the targeted subscore. Kelley's regression model uses the group average, OPI uses the IRT estimated person overall ability and Wainer et al.'s method uses an examinee's performance on other subsets of items. The intention of augmentation is to reduce the standard error (SE) of the estimated subscores, hence, to make the estimated subscores more reliable.

Reporting Subscores with Unidimensional Item Response Theory

The non-augmented methods to report subscores using UIRT models conduct separate item parameter calibration for each subset of items in a test and estimate latent subscores using calibrated item parameters. Another approach is to calibrate item parameters for all items in the test concurrently, then score examinees based on their responses to the subset of items assessing the same sub-domain. When conducting concurrent calibration, the item parameter estimation for one subset of items used ancillary information from other items in the test, therefore, each method is also an augmented method (Bock, Thissen and Zimowski, 1997).

Wainer et al.'s augmented subscore estimation with IRT theta score. Wainer et al. (2001) has proposed a method to estimate augmented subscores with

information from other subscores. The application of such a method to the IRT theta score is very similar to its use with the CTT-based subscore. The application of Wainer et al.'s (2001) method in UIRT is a multi-stage special case of a MIRT model, where each score is considered as a dimension. In other words, each item loads on only one subscale. Such structure is also referred to as independent clustering (Thissen & Edward, 2005). Wainer et al.'s subscore augmentation with IRT theta scores follows 4 stages—1) calibrating item parameters, 2) estimating IRT subscale theta scores, 3) calculating reliabilities at each theta level and the observed score covariance matrix of the IRT theta scores 4) regressing the estimated "true" IRT theta score on all subscale scores. The UIRT theta scores can be maximum likelihood estimates (MLE), maximum a posteriori (MAP) or expected a posteriori (EAP). When using MLE, the theta subscale estimates can be directly used in Wainer et al.'s augmentation as described in the CTT observed subscore reporting procedure. However, MAP and EAP are already reduced to the population mean in estimation (Thissen & Orlando, 2001). The amount of MAP and EAP shrinking to the population mean is related to the information of a given response pattern. Specifically, less information leads the MAP or EAP estimates to shrink more towards the mean, and vice versa. A correction needs to be made to use MAP and EAP in Wainer et al.'s augmentation (see in Equation 19). Assuming the SEs for all MAP estimates and EAP estimates are constant, the adjusted MAP or EAP is calculated as dividing the estimated MAP or EAP by the reliability of subscale v. The reliability of subscale v is denoted as ρ_v in Equation 19. The reliability of subscale v is calculated as the ratio of the variance of the MAP or EAP estimates for the subscale

to the sum of the variance of the MAP or EAP estimates and the average SE for MAP estimates or EAP estimates. The corrected MAP or EAP, denoted as $MAP(\theta_c)$ and $EAP(\theta_c)$ respectively in Equation 19 can be plugged into Equation 18 to compute the augmented IRT subscores.

$$MAP(\theta_c)\ or\ EAP(\theta_c) = \frac{MAP(\theta_{est})\ or\ EAP\ (\theta_{est})}{\rho_v}, \quad \text{where} \quad \rho_v = \frac{\sigma^2_{MAP(\theta_{est})\ or\ EAP\ (\theta_{est})}}{\sigma^2_{MAP(\theta_{est})\ or\ EAP\ (\theta_{est})} + \overline{\sigma_e^2}}. \tag{19}$$

Reporting Subscores with multidimensional Item Response Theory

MIRT models. MIRT models were developed as a realization of the complicated construct structure in assessment (Reckase, 2009). Like UIRT, MIRT models the probability of obtaining a correct answer to an item as a function of person's ability and item characteristics, such as item discrimination and item difficulty. The difference is that a MIRT model assumes that more than one underlying construct affects the examinee's performance to an item/test (Reckase, 2009, p. 59). Based on the relationship between latent traits and item responses, MIRT model can be compensatory or non-compensatory. In a compensatory model, an examinee's overall ability is modeled in the form of a weighted sum of dimensional abilities. Therefore, having a high ability in one dimension can compensate for deficiency on another dimension (Reckase,1997). As an example, a 2PL compensatory logistic MIRT model (Reckase, 1985; 1997) can be formulated as

$$P\left(u_{ij} = 1 \middle| \boldsymbol{a_i}, d_i, \boldsymbol{\theta_j}\right) = \frac{e^{(\boldsymbol{a_i}\boldsymbol{\theta_j} + d_i)}}{1 + e^{(\boldsymbol{a_i}\boldsymbol{\theta_j} + d_i)}}, \tag{20}$$

where the probability of getting item i correct is a function of d_i, the item difficulty related scalar, $\boldsymbol{a_i}$, a vector of item discrimination parameters, and $\boldsymbol{\theta_j}$, a vector of

abilities for person j. The exponent can also be written using parameters that are commonly used in UIRT model— $\sum_{k=1}^{n} a_{ik}\left(\theta_{jk} - b_{ik}\right)$, where a_{ik} and θ_{jk} are the k^{th}element of $\boldsymbol{a_i}$ and $\boldsymbol{\theta_j}$, respectively for an item with n dimensions; $d_i = -\sum_{k=1}^{n} a_{ik} b_{ik}$.

In a non-compensatory MIRT model, the overall probability is computed through a multiplication procedure, which requires an examinee to have high abilities (i.e., above certain levels for each dimension) on all dimensions to be able to answer the item correctly. In other words, having a low ability on one of the dimensions will necessarily have a negative impact on the probability of obtaining a correct answer regardless of abilities of other dimensions. The non-compensatory MIRT model was formulated by Sympson (1978) and Whitely (1980) as follows.

$$P\left(u_{ij} = 1 \middle| \theta_j\right) = c_i + (1 - c_i) \prod_{k=1}^{n} \frac{e^{1.7 a_{ik}(\theta_{jk} - b_{ik})}}{1 + e^{1.7 a_{ik}(\theta_{jk} - b_{ik})}} , \qquad (21)$$

where the probability of getting item i correct is calculated as the product of the logit for each dimension. Figure 2 presents the item characteristic surface for the 2PL compensatory logistic MIRT model (on the left) and the item characteristic surface of a 3PL non-compensatory IRT model (on the right). The lower asymptote for the 3PL non-compensatory IRT model is 0.068 meaning the probability of getting an item correct is above 0 when the examinee's ability is very low.

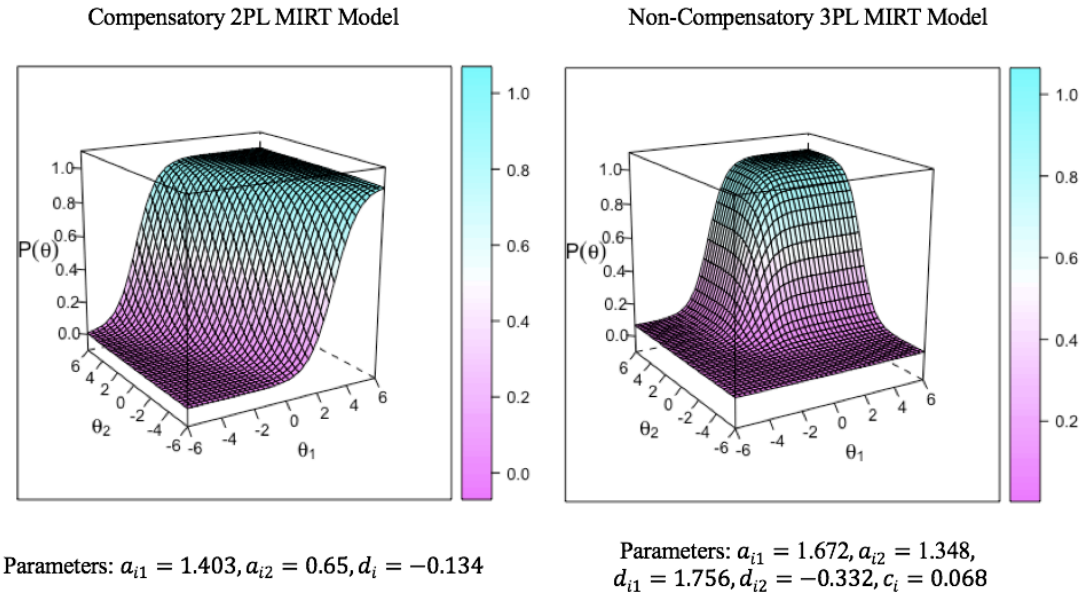


Figure 2. Item characteristic surfaces for a compensatory 2PL MIRT model and a non-compensatory 3PL MIRT model

According to Figure 2, in the compensatory 2PL model, the probability of getting an item correctly can be high when the ability on the first dimension, denoted as θ_1 in Figure 2, is high, and even when the ability on the second dimension, denoted as θ_2 in Figure 2, is relatively low. This means that the high ability on one dimension can compensate for the low ability on the other, whereas the probability of obtaining a correct answer in the non-compensatory model, requires high abilities on both dimensions.

Report subscores using MIRT models. As shown in Figure 1, latent traits are correlated in a multidimensional test with a simple or complex structure. Therefore, MIRT models are also augmented in that the estimation of one latent trait "borrows" information from other latent traits. Like augmentation in the CTT and UIRT framework, the augmentation in MIRT models increases measurement precision when estimating subscale abilities. Wang et al. (2004) conducted two

empirical data analyses, one with dichotomous responses to a science test with 5 subscales and the other with Likert-scale response data to a teacher personality inventory with 10 subscales. They compared the performance of the multidimensional approach (i.e. a multidimensional random coefficients multinomial logit model; Adam et al, 1997) and the unidimensional approach where subscores (i.e. subscale person abilities) are estimated based on subtests. Their results showed that the multidimensional approach improves the measurement precision in both analyses in terms of the subscore reliability and the number of items needed for achieving the same test reliability. The significant improvement of the measurement precision of the multidimensional approach comparing to the unidimensional approach happens when the length of the subtest is short, the correlation between subscales (i.e. dimensional latent trait) is high and the number of subscales is large (p.125). Similar results were also found in adaptive tests. Segall (1996) found that the multidimensional adaptive test achieved comparable measurement precision with approximately 30% fewer items than what were needed for a unidimensional adaptive test. Luecht (1996) also found that the multidimensional adaptive approach requires 25% to 40% fewer items than the unidimensional adaptive approach in a licensing context. In addition, de la Torre and Patz (2005) proposed the use of hierarchical Bayesian estimation for estimating parameters in a MIRT model (i.e. a 3PL compensatory MIRT model; Reckase, 1996). They conducted a simulation study to evaluate how the number of subscales/latent traits (i.e. 2 and 5), the length of subtests (i.e. 10, 30 and 50 items) and the correlation between subscales/latent traits (i.e. 0, 0.4, 0.7 and 0.9) influence on the precision of MIRT parameter estimates. Their study

compared hierarchical Bayesian subscore estimates with those yielded from the unidimensional approach in terms of the correlation between the estimated and the true abilities, and the ratio of the mean squared error (MSE) of the MIRT ability estimates to that of the UIRT ability estimates. The results indicated that the hierarchical Bayesian estimation of subscores using a MIRT model is more efficient across all conditions than the unidimensional EAP estimates which considers only one subscore at a time in scoring. Such advantage in estimation efficiency is larger when subtests are short and the correlation between subscales is high (de la Torre & Patz, 2005).

In addition to reporting subscores using a dichotomous MIRT model, polytomous MIRT models are also developed for estimating subscores for test items scored with partial credit scores. de la Torre (2008) developed a Generalized Partial Credit (GPC) MIRT model as an extension of the unidimensional GPC model (Muraki, 1992). A simulation study was conducted to evaluate the performance of the GPC MIRT model using the hierarchical Bayesian estimation by manipulating the number of score categories (i.e. 2, 3 and 4), the number of tests in the test battery (i.e. 2 and 5), the length of the (sub)test (i.e. 5,10 and 20 items) and the correlation between subscales/ abilities (i.e. 0, 0.4, 0.7 and 0.9). Similar to findings in de la Torre and Patz (2005), the accuracy of the subscale ability estimates from the GPC MIRT model is greatly improved when the correlation between latent traits is high. In addition, the stability of the estimates (assessed by posterior variance) is higher when there are more score categories, more items in a subtest, more subtests and higher correlation between subscales/latent traits. de la Torre (2008) also found that better

estimates of the correlation between subscales/latent traits could be obtained with longer tests and higher correlation. The use of the GPC MIRT model is also demonstrated with a real data analysis from a multidimensional test battery with math subtests and science subtests.

Yao and Boughton (2007) used the Bayesian estimation in BMIRT to estimate subscores (i.e. subscale scores) with a 3PL compensatory MIRT model (Reckase, 1997) and a 2PL partial credit MIRT model (Yao & Schwarz, 2006) for a mixed format test. In this study, the authors manipulated sample size (i.e. 1000, 3000 and 6000 examinees) and the correlation between subscales/latent traits (i.e. 0, 0.1, 0.3, 0.5, 0.7 and 0.9). The Bayesian estimates yielded from the BMIRT software is compared with (a) the percentage correct on number-correct subscores, (b) the OPI subscores (Yen, 1987), (c) the maximum likelihood MIRT subscores and (d) the Bayesian UIRT subscores with the criteria of the subscore estimation accuracy and the classification accuracy. Yao and Boughton (2007) have concluded that (a) the BMIRT estimation is the most accurate method in terms of both subscore recovery and classification accuracy rates, (b) the maximum likelihood MIRT and UIRT subscores are comparable to that of the BMIRT using Bayesian estimation when the correlation between subscales/latent traits are low, (c) the OPI, as an augmented method, yielded similar results to the BMIRT estimates in terms of both ability parameter recovery and classification accuracy when correlations are high.

Studies mentioned above can only report subscores for a test. de la Torre and Song (2009) proposed the higher-order IRT (HO-IRT) model to estimate the summative score (i.e. overall ability) and the subscore (i.e. domain abilities) at the

same time. A diagram is presented in Figure 3 to demonstrate the model structure, θ_1, θ_2 and θ_3 are domain-specific abilities (i.e. subscores); θ_g is a single higher-order ability or general ability (i.e. the summative score) to account for the correlation between the domain-specific abilities. The rate of the overall ability attributable to a specific domain score, denoted as λ_{θ_d}, is fixed for all examinees. In the context of MIRT, λ_{θ_d} and a_{id} are normally constrained to be positive to indicate positive correlations between the domain-specific ability (i.e. subscore) and the general ability (i.e. summative score), and those between item score and domain specific ability (i.e. subscore), respectively. When there are only two subdomains in a test, λ_{θ_1} is constrained to be equal to λ_{θ_2} to avoid scale indeterminacy. de la Torre and Song (2009) conducted a simulation study to examine the performance of the HO-IRT model where they manipulated the number of subtests (i.e. 2 and 5), the number of items in each subtest (i.e. 10, 20 and 30), the correlation between subscale/ subdomain abilities (i.e. 0, 0.4, 0.7 and 0.9) and sample size (i.e. 1000, 2000 and 4000). The subscores and summative scores yielded from the HO-IRT model were compared with the estimates based on a conventional UIRT (CU-IRT) approach. Judging by the variance of the posterior distribution, the summative score estimates yielded from the CU-IRT approach is more precise than that yielded from the HO-IRT model across all levels of latent trait correlation. In terms of the subscores, the improvement of HO-IRT estimates from the CU-IRT estimates is negligible when the correlation between subscales\latent traits is low and when test is long enough. When there are more subtests and the correlations between the subscales\latent traits are

high, the HO-IRT model estimates are found to be more precise than the estimates yielded from the CU-IRT model.

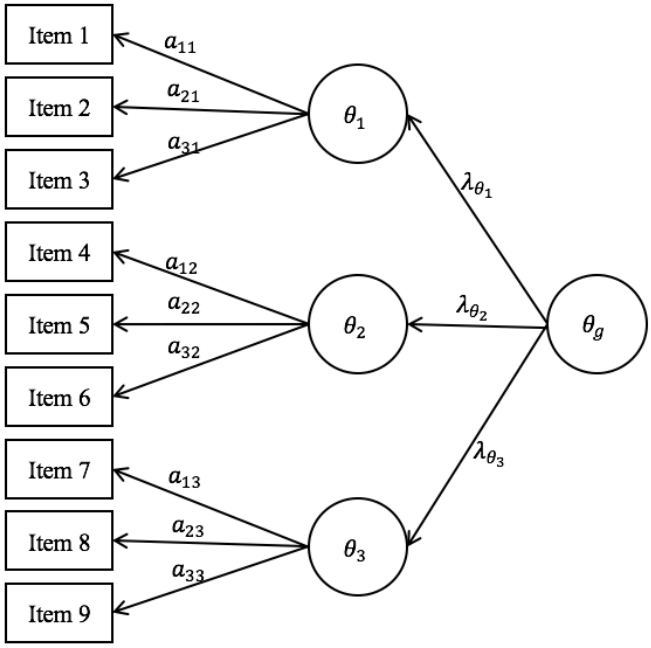


Figure 3. Factor structure of HO-IRT model (de la Torre & Song, 2009)

The bi-factor model (Gibbons & Hedeker, 1992) as a special case of hierarchical and higher-order factor model typically assumes orthogonality between the content domains related factors. Only a couple of studies investigating the use of the bi-factor MIRT model in subscore reporting were found in the literature. DeMars (2005) compared a bi-factor MIRT model, a two-factor model (i.e. simple structure MIRT model), Wainer et al.'s (2001) augmentation on UIRT estimates and unaugmented UIRT estimates in a simulation study based on parameters estimated from real datasets from two tests. Results indicated that the unaugmented UIRT method has the largest bias and root mean squared error (RMSE), two MIRT methods yielded comparable bias and RMSE, on one test higher than that of the augmented UIRT estimates, on the other test lower. It is concluded that there is no clear winner

among the bi-factor method, two-factor model and the augmented UIRT approach, but they are all preferred over unaugmented UIRT method. Md Desa (2012) proposed a bi-factor confirmatory compensatory model and a bi-factor confirmatory partially compensatory MIRT models to enhance subscore reliability and classification accuracy with the Bayesian estimation.

Reporting Subscores for Tests with Complex Structure. Very few studies have discussed reporting subscores for tests with complex structure. Boughton et al. (2006) investigated the use of a MIRT approach to report subscores for a test with complex structure. In the simulation study, they have manipulated sample size, the correlation between subscales, the number of items contributing to each subscale and the number of items with complex structure. Their study concluded that as the correlation between latent traits increased, the error in parameter estimation also increased for all subscales with complex structure items, yet that decreased for subscales consisting of only simple structure items.

Feinberg and Wainer (2014) conducted a simulation study to investigate under what circumstances MIRT estimated subscores have added-value to the holistic test score. They manipulated the number of unique items in a subscale (i.e. 5, 10, 20, 30 40 and 50), the correlation between subscales (i.e. 0, 0.3, 0.7 and 0.9), the percentages of overlapping items (i.e. 11 levels ranging from 0% to 100%, evenly spaced) and their loadings on each subscale (i.e. complex structure—both loadings drawn from the same distribution as the loadings for the unique items; semi-complex structure—one loading drawn from the same distribution as for unique items, and the other drawn from a distribution with a smaller mean and standard deviation, and moderate

structure— both loadings drawn from a distribution with a smaller mean and standard deviation). The estimated MIRT subscores (i.e. domain abilities) are assessed by (a) reliability and (2) score orthogonality. They concluded that the removal of overlapping items will always improve the added value of the subscores, though the inclusion of such items increases the reliability of the subscores.

Both of these studies were conducted with complex structure items loading on two construct dimensions. In many cases, subscores are designed based on content specification in a test blue print rather than the construct multidimensional structure. The MIRT model outlines a structure where a cluster of items loads on more than one latent traits. Since latent traits are not observable, they represent constructs. However, when subscores are designed based on content specification, items assessing the same content domain may not load on the same latent traits as what MIRT has suggested. In fact, the test itself may be unidimensional (e.g., the NAEP math test). Therefore, the use of MIRT models for reporting content-based subscores is questionable. In this study, a double-coded item refers to an item that measures knowledge from two sub-content domains. Since an item does not necessarily measure two latent traits (i.e. construct dimensions) when it tests two sub-content domains, a UIRT approach is adopted. Subscores are reported based on item difficulty decomposition for different content domains.

Method Comparison and Summary

Current methods for reporting subscores or estimating subscale abilities have been reviewed in the previous sections. To emphasize the advantages and/or disadvantages of these methods, this section synthesizes studies comparing these

methods. This synthesis is based on the comparison framework in Longabach (2015) where comparisons are categorized into CTT methods vs. IRT methods, MIRT methods vs. UIRT methods, and unaugmented methods vs. augmented methods.

CTT vs. IRT methods. Unaugmented CTT subscores in comparison studies, including the standardized number of correct scores in Luecht (2003), the percent correct on subscales (Dwyer, Boughton, Yao, Steffen, & Lewis, 2006; Yao & Boughton, 2007) and the proportion correct score in Shin (2007), are all variations of the number-correct raw subscores. They generally performed worse than augmented CTT methods, IRT methods and MIRT methods in terms of parameter estimation accuracy (Dwyer et al., 2006; Luecht, 2003; Yao & Boughton, 2007) and reliability (Haberman & Sinharay, 2010) Luecht, 2003; Shin, 2007). For augmented CTT subscores, Yen's OPI (1987) performs as well as MIRT models when the subscales/latent traits correlated strongly (Yao & Boughton, 2007); Wainer et al.'s (2001) CTT-based augmentation yielded better reliability for subscores than OPI (Shin, 2007). DeMars (2005) found that the unaugmented UIRT method had the largest bias and RMSE when compared with the MIRT approaches and Wainer et al.'s augmentation of the UIRT theta scores (2001). Thissen and Orlando (2001) concluded that the IRT scoring method produces scores that are linearly related to the underlying latent traits, which makes it more useful than the sum scores or the number-correct raw scores when investigating the relationship between test scores and external variables. General advantages of IRT methods including flexibility in scaling and high precision in scoring have been summarized when introducing the IRT model. One thing should be kept in mind is that the key premise for IRT methods

producing more accurate ability estimates and better reliability in subscore reporting is the data meets the assumptions for the IRT model.

UIRT vs. MIRT methods. A number of comparison studies between the UIRT and MIRT approaches were conducted when proposing new estimation methods for MIRT models (de la Torre & Patz, 2005; Wang et al., 2004; Yao & Boughton, 2007) or when proposing newly-developed MIRT models (de la Torre, 2008; de la Torre & Song, 2009). Most of these studies were reviewed in detail previously when discussing applications of MIRT models in subscore reporting. The findings of these studies are largely consistent. These studies found that the MIRT approaches yielded more precise and reliable estimates when (a) the correlation between subscales/latent traits are high (Boughton, et al., 2006; de la Torre & Patz, 2005; Wang et al., 2004; Yao & Boughton, 2007), (b) there are more subtests/subscales (Wang et al., 2004; de la Torre & Song, 2009) and (c) the subtest is short (de la Torre, 2008; de la Torre & Patz, 2005; de la Torre, Song, & Hong, 2011; Wang et al., 2004). In the study by de la Torre and Song (2009), the HO-IRT method was found to perform better than the CU-IRT method in optimal conditions. However, when evaluating the overall ability estimates yielded from the HO-IRT and the CU-IRT, results showed that the correlation between the true and estimated abilities are nearly identical for both methods when the subdomain abilities are correlated, and the CU-IRT estimates have smaller posterior variance across all levels of correlation between latent traits, hence is more precise.

Some comparison studies concur with de la Torre and Song (2009) in that the MIRT approaches produce similar estimates with UIRT approaches for estimating

subscores (e.g., DeMars, 2005; de la Torre & Song, 2009; Dwyer et al., 2006; Gessaroli, 2004). For example, Gessaroli (2004) and Dwyer et al. (2006) both found that MIRT approach yielded similar estimates with the augmented methods involving correlational structure, such as Wainer et al.'s augmentation with UIRT theta scores. Further, some studies pointed out that MIRT approaches produce estimates that are not so different from UIRT estimates while being computationally challenging (Longabach, 2015). For instance, Luecht (2003) compared the number of correct score, UIRT estimates with item parameters calibrated using subtest data (UIRT-S), UIRT estimates with item parameters calibrated using the total test data (UIRT-T) and MIRT estimates. The UIRT-T approach was selected after the analysis for its calibration efficiency. The MIRT model was not selected because "the complexity of using a multidimensional model is hard to be justified" (Luecht, 2003, p. 14). Haberman and Sinharay (2010) have also suggested that testing programs with limited time for data analysis may not favor MIRT. In addition to the computational complexity, the interpretation of MIRT estimates is challenging.

Unaugmented vs. Augmented Methods. Augmented methods for subscore reporting referred to the methods that incorporate information from sources other than the target subscore. Kelley's regression method used information from the group mean, Yen's OPI (1987) stabilizes estimates of objective abilities with total scores and Wainer et al.'s (2001) augmentation "borrows" information from other subscores. In addition, estimating subscores using UIRT item parameters calibrated with data from the whole test is also an augmentation method (Bock et al., 1997). When calibrate item parameters using all data, the collateral test information is used since

responses to items in other objective scores also contributed to the parameter estimation of items that belong to the targeted domain score. As introduced, the MIRT model is also an augmentation, as the correlation between latent traits allowed information from subscales to be utilized when estimating subscale scores. Studies have shown that the augmentation methods improved the subscore estimation precision and reliability (e.g., Dwyer et al, 2006; Wainer et al., 2001; Puhan et al., 2010; Sinharay, 2010; Skorupski & Carvajal, 2010; Skorupski, 2008). Yet, augmented subscores can be highly correlated in many situations. For example, Stone, Ye, Zhu and Lane (2009) conducted a real data analysis with data of a large-scale mathematics test. They compared Yen's OPI (1987), Wainer et al.'s (2001) augmentation, MIRT methods and unaugmented UIRT. Their study found that all three augmented methods yielded more precise subscores than the unaugmented UIRT method. Before estimating subscores, Stone et al. (2009) conducted an exploratory factor analysis and determined that the test is unidimensional. Therefore, subscores were designed in terms of content clustering. Under this situation, they found that the subscores are highly correlated for all three augmented methods (e.g., the subscores based on Wainer et al.'s (2001) method are almost perfectly correlated), whereas the correlations for the unaugmented IRT subscores are much lower.

Skorupski (2008), Stone et al. (2009) and de la Torre et al. (2011) have all stated that the purpose of reporting subscores should determine the methods used to estimate the subscores. For example, when subscores are used primarily as diagnostic information which informs future learning and instruction, MIRT approaches can be appropriate as it incorporates as much information as possible and considers the

relation between subscale performance (de la Torre & Patz, 2005). However, when subscores are used to make high-stakes decisions such as graduation or admission, the justification of augmented subscores are limited by the complexity in the meaning of the score. Therefore, reliable subscores without integrating any ancillary information would be more appropriate to use for a high-stake test (Longabach, 2015).

In the current study, the proposed model is based on the UIRT framework where only one underlying latent trait is assumed to have an impact on the item responses. The adoption of the UIRT framework echoes the goal of reporting subscores formed for content dimensionality. In addition, the intended application of the proposed model is to a single test, rather than a test battery. Like Stone et al. (2009) have discussed, content clustering in a test is difficult to be justified as construct multidimensionality. When content dimensionality underlies the subscores, it makes sense to use UIRT where the difference among subscores for different content domains within an examinee is only attributable to the item characteristics for items in that content, as indicators of content difficulty; the difference across examinees for the same subscore is only attributable to examinees' latent abilities.

As it follows the UIRT framework, the proposed model is computationally less challenging than the MIRT model. At the same time, the meaning of the subscores is clearer in interpretation.

To model responses to the double-coded items, reporting subscores with the proposed model requires all item parameters to be calibrated in a single calibration using item response data from the whole test. Hence, the proposed model is augmented in a way that is very similar with that described in Bock et al. (1997).

Only the item parameters related to the reported subscore in a double-coded item will be used for scoring. The problem of obtaining highly correlated augmented subscores is not the key concern in this study. As the model to report subscores for sub-content domains and the test is assumed to be unidimensional, the correlation between subscores represents how knowledge/contents covered in each of the subscores relate rather than being a representation of examinees' abilities on different constructs.

The Model with Restrictions on Item Difficulty (MIRID)

This section summarizes literature on the MIRID. This synthesis on the MIRID is laying a theoretical basis for the proposal of the 2PL-DT-MIRID. To do so, this section first introduces the background of the MIRID model, outlines model formulation, discusses the relationship between the logistic linear test model (LLTM; Fischer, 1973, 1983) and the MIRID, describes relevant extensions based on the dichotomous Rasch MIRID and finally summarizes the estimation methods and software.

Background of the Model with Restrictions on Item Difficulty (MIRID)

A traditional IRT modeling approach considers the relationship between ability/latent trait on a construct and item responses as summative. Such a modeling approach ignores the cognitive or behavior process underlying the causal sequence of item responses. Since the advancement in cognitive psychology and promotion of reporting diagnostic scores in educational assessment, componential models were developed to recognize intermediate cognitive processes as well as the final item

responses and to explain final responses based on intermediate responses (e.g. Butter et al. 1998; Huang, 2011; Li, 2017).

In a descriptive framework, the person ability in the Rasch model can be considered as an intercept and the item indicator as a predictor, so that each item has a specific effect (Wilson & De Boeck, 2004). Built upon the descriptive perspective, the LLTM (Fischer, 1973, 1983) was developed as an explanatory model, where the item properties are modeled as the predictors. The LLTM models the item difficulty as a linear composite of component difficulties and the component weights (Fischer, 1973, 1983). In LLTM, the item parameters to be estimated are the component difficulty and the intercept in the linear combination. The component weights, that is, to what extent the component difficulty contributes to the linear composite of item difficulty is known. Such a formulation provides a more parsimonious approach to modeling item effects, but imposes additional assumptions (Wilson & De Boeck, 2004).

The MIRID model was originally proposed by Butter et al. (1998) based on Butter (1994) and De Boeck (1991). The MIRID is proposed to model item responses to an item family that consists of a number of component items (i.e. subtasks) and a composite item (i.e. composite task). Essentially, the goal of MIRID is to investigate how different components impact on one's performance on the composite task by decomposing the composite item difficulty into a linear combination of component difficulties. In the MIRID, the values of component weights are no longer assumed to be known a priori. Instead, the component weights, the component difficulties and the intercept are all to be estimated. This is made possible by modeling responses to

subtasks along with the composite task. In other words, by estimating difficulties for component items with item responses to subtasks, the composite difficulty is expressed in terms of component difficulties. With the inclusion of component items, the MIRID can be applied to model psychological constructs and cognitive constructs in educational assessment. In the psychological measurement setting, for example, Lee (2011) demonstrated the use of the MIRID and its extensions to measure the complexity of guilt, based on subtasks on norm violation, worrying and restitution. Butter et al. (1988) presented an application of MIRID in an educational measurement setting with a spelling test where a student needs to master two rules to be able to produce correct spelling of the plural form of a given noun. In this example, the MIRID showed its great capability of providing diagnostics information for learning.

Rasch MIRID Model

The MIRID was proposed as a Rasch family model for dichotomous response data (Butter et al., 1988). As introduced previously, two types of items are needed to use the MIRID—the component item and the composite item. A composite item contains tasks that can be decomposed to different kinds of subtasks. A component item can be an item that contains a generic subtask result from the composite task decomposition and it is specifically related to the composite task under study (Butter et al, 1998), or it can be a single operation (Li, 2017). Imagine a hypothetical arithmetic test that consists of arithmetic problems on addition and subtraction. A composite item is a problem measuring both arithmetic operations, the two component items for this composite item should be one on addition and the other on

subtraction. Table 1 presents the structure of such a test in the case where there are two composite items. In Table 1, a "1" indicates that the arithmetic problem involves that component, a "0" indicates that the arithmetic problem does not contain that component. Items 1 and 2 are component items of composite item 3; items 4 and 5 are components of composite item 6. In measuring a psychological construct, items in an item family can also be nested within a situation.

Table 1
An Arithmetic Test with Two Item Families

		Component 1 (Addition)	Component 2 (Subtraction)
Item Family 1	Item 1	1	0
	Item 2	0	1
	Item 3	1	1
Item Family2	Item 4	0	1
	Item 5	1	0
	Item 6	1	1

For a test with I item families each with K components taken by J examinees, the probability of person j $(j = 1 \dots J)$ answering the $k^{th} (k = 1 \dots K)$ component in the i^{th} $(i = 1 \dots I)$ item family correctly, denoted as $P\left(X_{jik} = 1 \middle| \theta_i\right)$, can be formulated as follows.

$$P\left(X_{jik} = 1 \middle| \theta_j\right) = \frac{\exp\left(\theta_j - \beta_{ik}\right)}{1 + \exp\left(\theta_j - \beta_{ik}\right)}, \tag{22}$$

where the person ability is denoted as θ_i; the component item difficulty is denoted as β_{ik} for component k in item family i.

For a composite item, $k = 0$. The probability of answering a composite in item family i correctly can be modeled as

$$P\left(X_{jik} = 1 \middle| \theta_i\right) = \frac{\exp\left(\theta_i - \sum_{k=1}^{K} \sigma_k \beta_{ik} - \tau\right)}{1 + \exp\left(\theta_i - \sum_{k=1}^{K} \sigma_k \beta_{ik} - \tau\right)}, \tag{23}$$

where the item difficulty of the composite task is a weighted sum of the component difficulty, denoted as $\sum_{k=1}^{K} \sigma_k \beta_{ik}$, and a scaling intercept, denoted as τ. In Equation 23, component difficulty β_{ik} is weighted by σ_k for component k.

For the arithmetic test in the example presented previously, the item family i ($i = 1, 2$) with 2 component items and one composite item, the item difficulties for the six tasks can also be represented in the matrix form as

$$\begin{pmatrix} \delta_{10} \\ \delta_{11} \\ \delta_{12} \\ \delta_{20} \\ \delta_{21} \\ \delta_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1 & \sigma_2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1 & \sigma_2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \\ \tau \end{pmatrix}, \tag{24}$$

where $\delta_{ik} \equiv \beta_{ik}$ ($i = 1 \ldots I, k = 1 \ldots K$) and $\delta_{i0} \equiv \beta_{i0}$. Note that the model is not identified in this example. The degree of freedom for the Rasch model is $I(K + 1) - 1$, and degree of freedom of the MIRID is $IK + K$. As a dichotomous Rasch MIRID is a restricted Rasch model, the MIRID is only identified when the degree of freedom of the MIRID is larger than that of the Rasch model. That is, $I - K - 1 > 0$. When $I - K - 1 = 0$, MIRID is simply another parameterization of the Rasch model. In the case where $I - K - 1 < 0$, the model is overparameterized, therefore, not identified. In the example of two item families each with two component items and one composite item, $I - K - 1 = -1$. The Model is not identified. The model would be identified if more item families were added to the test.

MIRID vs. LLTM

Since research has indicated that LLTM and MIRID are similar and equivalent with the same model constraints (Bechger, Verstralen, & Verhelst, 2002;

Butter et al., 1998; Maris & Bechger, 2004), it is critical to compare and contrast these models. As described previously, the LLTM is a componential model where the item properties are used to explain the difference between items with respect to their impact on the probability of obtaining a correct answer for an item (Wilson & De Boeck, 2004). The reason that the LLTM is considered as an explanatory model is that the item difficulty in the Rasch model is decomposed to a weighted sum of the item properties. In other words, the item difficulty for an item has been explained by the decomposition. The item difficulty decomposition in the LLTM can be represented as in Equation 25.

$$\delta_i' = \sum_{k=0}^{K} \beta_k q_{ik} \,, \tag{25}$$

where q_{ik} is the value of item i on property k, β_k represents the component difficulty for property k. When $k = 0$, with $q_{i0} = 1$, β_0 is the item intercept.

There are a few things that the MIRID and LLTM share in common. First, both models decompose item difficulty into a linear composite. Second, in the linear composite for the composite item difficulty for MIRID and the item difficulty in LLTM, no error term has been included. This means that both models assume that the prediction of the item difficulty is perfect.

Contrasting the LLTM and the MIRID, they have three major differences. First, q_{ik} is a constant in LLTM. That is, the Q-matrix for I item and K components is known a priori. In the MIRID, all parameters in the linear composite item difficulty, including, β_{ik}, σ_k and τ, are estimated. Second, since the LLTM uses a known Q-matrix, it does not need component items in the test. On the other hand, the MIRID requires component items to estimate composite item difficulty. Third, in the LLTM,

q_{ik} is an indicator showing if the k^{th} item property is needed. In the case where q_{ik} is not restricted to take on the value of 0 or 1, it indicates how many times or to what extent the item property is needed (Butter et al., 1998). The weights are item and component specific. Whereas in the MIRID, the weight σ_k is not item specific, which means that the component difficulty for component k is weighted the same for all composite items that involve component k. But the component difficulty for component k can be different in different item families.

Since the goal of the proposed model is to report subscores for a test with double-coded items, the key part in modeling is to understand and investigate to what extent the content knowledge in each domain is needed in solving the composite question. In reality, the Q-matrix in the LLTM is unknown. Specifically, which components contribute to the composite is known, but the extent to which the component difficulty contributes to the composite difficulty is unknown. Therefore, the LLTM is not a good candidate for modeling double-coded items in the current study. On the other hand, the MIRID estimates both the component weights and the item-family specific component difficulties (Note that the property/component difficulty in LLTM is not item specific.). The structure of MIRID enables the same component to have different component item difficulties in different component items. In the example of an arithmetic test on addition and subtraction, assuming item family A measures one-digit addition and subtraction, and item family B measures two-digit addition and subtraction. Two components, one in A and one in B, both measure addition, and their item difficulty would be expected to be different.

Therefore, the MIRID, rather than LLTM is selected as the basis for modeling double-coded items in the current study.

Extensions of the MIRID

After the initial proposal of the binary Rasch MIRID (Butter et al., 1998), extensions of the MIRID have been developed to generalize the use of the MIRID to a broader range of testing data. For example, the Rasch MIRID was extended to model polytomous data based on the graded response model (Samejima, 1968) and the partial credit model (Masters, 1982) using cumulative logit and adjacent-category logit, respectively (Lee & Wilson, 2009; Wang & Jin, 2010b). A multilevel, two-parameter, random weight extension was proposed to model (a) ability variation with level-2 predictors, (b) item-component interaction by using random weights and (c) item discrimination power difference among items by incorporating item discrimination parameters (Wang & Jin, 2010a); Lee (2011) and Lee and Wilson (2017) have generalized the MIRID with random item effect and multidimensionality; Li (2017) has proposed a model based approach to detect differential item functioning with the MIRID.

Since the extensions of the MIRID are not the focus of this study, they will not be described in a great detail. The proposed model is a two-parameter doubly-testlet MIRID that is based on the level-1 model of the extended MIRID proposed by Wang and Jin (2010a) where the item discrimination or the slope parameter is added into the original binary Rasch MIRID. The level-1 model for a composite item of Wang and Jin (2010a) is presented in a consistent form with the Rasch MIRID (Equation 24) in Equation 26.

$$P(X_{jik}=1|\theta_i)=\frac{\exp(\alpha_{ik}\theta_i-\sum_{k=1}^{K}\sigma_k\beta'_{ik}-\tau')}{1+\exp\left(\alpha_{ik}\theta_i-\sum_{k=1}^{K}\sigma_k\beta'_{ik}-\tau'\right)}, \quad (26)$$

where α_{ik} is the slope parameter for component k in item family i, β'_{ik} is the item difficulty parameter applied to the $\alpha_{ik}\theta_i$ scale. An alternative way to formulate Wang and Jin (2010a)'s level-1 model in a standard IRT representation is as follows.

$$P(X_{jik}=1|\theta_i)=\frac{\exp(a_{\mathrm{ik}}(\theta_i-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau))}{1+\exp\left(a_{\mathrm{ik}}(\theta_i-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau)\right)}, \quad (27)$$

where $a_{ik}=\alpha_{ik}$, $\beta_{ik}=\frac{\beta'_{ik}}{\alpha_{ik}}$ and $\tau=\frac{\tau'}{\alpha_{ik}}$.

The reason for choosing the two-parameter MIRID for the proposed model is to demonstrate its capability of reporting subscores using IRT-based NCS. The IRT-based NCS is to score examinees' performance using a test characteristic curve (TCC) based on the item parameters calibrated. The scoring proceeds with finding the point on the theta scale that corresponds to the number-correct score on TCC. Thus, latent ability scores obtained from IRT-based NCS are the same for examinees with the same sum score. Since the number correct score is the sufficient statistic for the Rasch model, the IRT-based NCS method will yield ability estimates that are identical to those of the IRT pattern scores in the Rasch model. The IRT-based NCS scores are only different from the IRT pattern scores when item discrimination parameters differ for different items. To demonstrate the use of the IRT-based NCS, the two-parameter MIRID was selected for the proposed model.

The IRT-based number-correct scores are favored by some testing programs because in mapping the sum scores to the logit scale using TCC, IRT-based number-correct scores can be compared directly across test forms. In other words, the test-dependent sum scores are transformed to test-independent logit scores in the IRT-

based NCS. In addition, the IRT-based number-correct scores are easier to interpret than the IRT scores.

Model Estimation

Butter et al. (1988) described a likelihood-based approach to estimate parameters in the MIRID. Specifically, the conditional maximum likelihood estimation (CMLE) method was used when proposing the MIRID. When the component difficulty, β_{ik}, and the component weights, σ_k are both estimated, the model is not linear anymore (Maris & Bechger, 2004), and the likelihood conditioning on score $s_i (s_i = \sum_{j=1}^{J} \sum_{k=0}^{K} X_{ijk})$ is not an exponential family likelihood (Anderson, 1980). The first partial derivatives and the second partial derivatives are derived in Butter et al. (1988) based on the conditional likelihood function. The Newton-Raphson estimation was used to estimate the parameters interactively.

When developing extensions of the Rasch MIRID, the marginal maximum likelihood estimation (MMLE) is used to estimate parameters in some of the extensions of the MIRID (e.g. Wang & Jin, 2010a, 2010b; Smits & De Boeck, 2003). In general, MMLE is to integrate out the latent trait and estimate the item parameters using derivatives. Person parameters are then estimated based on the item parameters. Li (2017) described detailed procedures of estimating parameters in a Rasch MIRID with MMLE (p. 36-38). The Markov chain Monte Carlo (MCMC) was used in estimating parameters in more complex extensions of the MIRID model, such as the random item MIRID (Lee, 2011; Lee & Wilson, 2017), the multidimensional extension of the MIRID (Lee, 2011), and the multilevel cross-classified random effect MIRID (Huang, 2011).

Computer software. After Butter et al. (1988) proposed the dichotomous Rasch MIRID, Smits, De Boeck, & Verhelst (2003) developed the MIRID CML to implement CMLE for estimating parameters in the Rasch MIRID and One Parameter Logistic Model (OPLM) MIRID. The MIRID CML program uses the CMLE approach and the Davidson-Fletcher-Powell technique (Bunday, 1984) or the Newton-Raphson optimization technique (Bunday, 1984; Gill, Murray, & Wright, 1981) to estimate item parameters and their SEs. Person parameters and their SEs were obtained by a weighted maximum likelihood (Warm, 1989), subsequently. Smits et al. (2003) compared the estimates yielded from the MIRID CML and those from the SAS NLMIXED which uses MMLE. They found that the item parameter estimates are comparable, yet there were some differences in person parameter estimates. Smits et al. (2003) concluded that the MIRID CML supplemented with weighted maximum likelihood (Warm, 1989) should be preferred when individual ability estimates are required. The MIRID CML is less time consuming than the SAS NLMIXED, but the SAS NLMIXED is more flexible in estimating other extensions of the MIRID. Detailed instruction of using the MIRID CML and SAS NLMIXED to estimate parameters in the MIRID and in its extensions can be found in Smits, et al. (2003), Simts and De Boeck (2003) and Wang and Jin (2010a, 2010b).

Lee (2011), Lee and Wilson (2012) and Huang (2011) used MCMC algorithms to estimate parameters in complex MIRID extensions. They all used WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). Lee (2011) used the R2WinBUGS (Sturtz, Ligges, & Gelman, 2005) to achieve efficacy in summarizing the results in simulation studies.

Since the Rasch MIRID is a restricted version of the Rasch model, the structures of the Rasch MIRID and its straightforward extensions are similar to standard IRT models in many ways. Likelihood methods are first adopted. As the extension gets more complex, the Bayesian estimation is more preferred since it is a modern computer-intensive technique that simplifies the parameter estimation problem (Baker, 1998; De Boeck & Wilson, 2004; Gelman, Carlin, Stern, & Rubin, 2004; Tanner, 1996; Zeger & Karim, 1991).

Summary

This chapter reviewed literature on (a) IRT and TRT, (b) methods on subscore reporting and (c) the MIRID. The review of literature on IRT and TRT demonstrated the advantages of using modern measurement theory in modeling item response data, discussed the impact of item clustering effect resulting from using common stimuli and presented methods developed to partitioning the testlet effect and the latent trait, especially the TRT model for paired stimuli. Further, the methods for subscore reporting are also reviewed and compared. While summarizing the subscore reporting methods, limitations of the current methods in estimating content-based subscores for tests with innovative item types are presented. Subsequently, the literature on the MIRID was synthesized to justify the usage of the two-parameter extension of the MIRID for reporting content-based subscores for a test with double-coded items.

In the following chapter, the 2PL-DT-MIRID is presented. Following the introduction of the proposed model, a simulation study was designed to evaluate the performance of the proposed model across various optimal and sub-optimal conditions by manipulating factors that may have an impact on the accuracy and the

reliability of the estimated subscores. In addition, the procedures of reporting subscores using the IRT-based NCS using the proposed model is also demonstrated using the simulated datasets in Chapter 3.

Chapter 3: Method

In this chapter, the proposed model is first specified. A Monte Carlo simulation study is conducted to (a) evaluate the performance of the proposed model in terms of recovery of true model parameters, estimation of sub-content domain scores and score reliability across various study conditions, (b) compare the proposed model with other models that ignore the innovative item types in modeling item responses, and (3) highlight the use of the proposed model in IRT-based NCS. This chapter introduces the detailed technical procedures of the simulation study, including simulation conditions, data generation, model formulation (including all models compared in this study), model identification, model parameter estimation and model evaluation criteria.

A Non-Compensatory Two-Parameter Doubly Testlet MIRID

Double-coded items and scenario-based testlets are prevalent in authentic assessment of higher-order cognitive skills in large-scale assessment, such as SAT® and PARCC. In addition, since many subscores are designed according to content clustering in test specification rather than construct multidimensionality, the current methods for reporting subscores are limited for estimating sub-content domain scores for tests with double-coded items embedded in multiple context. The MIRID, on the other hand, carries out a UIRT approach where it assumes only one latent construct underlies an examinee's performance on an item. Further, the decomposition of item difficulty of a composite item in the MIRID sets up an applicable basis for (a) estimating content-based subscores, and (b) modeling item parameters for double-

coded items. Moreover, the MIRID model assumes that each composite item is the combination of corresponding component items. Such requirement coincides with the current educational assessment practice where content or skills are measured individually and jointly at different levels of difficulty. An example of an item family for arithmetic operation is presented as follows to demonstrate the structure of the component and composite item.

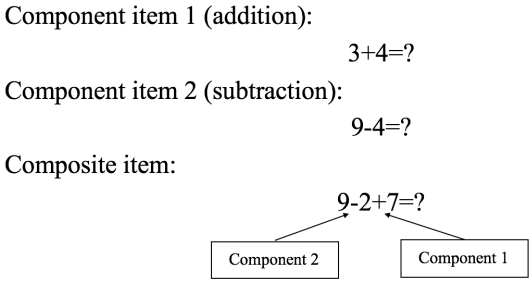


Figure 4 Example of an item family in the MIRID

In this example, one component item measures addition, the other measures subtraction. The composite item assesses the combination of addition and subtraction but with different numbers. With the benefit of the MIRID and the doubly testlet model developed by Jiao and Lissitz (2014), this study proposes a non-compensatory 2PL-DT-MIRID. The proposed model can estimate sub-content domain scores using both pattern scoring and IRT-based NCS.

The proposed non-compensatory 2PL-DT-MIRID for a double-coded item based on information from 2 testlets d_1 and d_2 is formulated as follows.

$$\mathrm{P}(\mathrm{X_{ji0}}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)} * \frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2}(i)}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2}(i)}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}. \quad (28)$$

For double-coded item i $(i = 1..I)$ with K components embedded in the context of two testlets, the probability of person j $(j = 1 \ldots J)$ answering composite item i $(i = 1 \ldots I)$ correctly is denoted as $P(X_{ji0} = 1|\theta_j)$. The item specific discrimination parameter is denoted as a_{i0}, item difficulty is decomposed into a weighted sum of the component item difficulty, denoted as β_{ik} for component k . The component weights are denoted as σ_k, and the composite intercept is denoted as τ. The model assumes that an examinee needs to master content in both testlets to be able to obtain a high probability of answering this item correctly. In other words, mastering one testlet does not compensate for a non-mastery or partial mastery of the other testlet. The non-compensatory relation is modeled by taking the product of the probability of answering such an item correctly considering only one of the testlet effects, denoted as $\gamma_{jd_{1(i)}}$, and that considering only the other testlet effect, denoted as $\gamma_{jd_{2(i)}}$. The non-compensatory relation is particularly true for items nested within paired-testlets since the items are designed to require knowledge from both testlets to assess students' ability to synthesize information and apply the synthetic knowledge to solving problems. Extensions of the proposed model can model item responses from items that measure more than two content domains (e.g. triple-coded items) or items based on more than two testlets. Yet, these extensions will not be evaluated in the current study. The current study focuses on evaluating the performance of the proposed model to pave out a basic functional form for future extensions.

Simulation Conditions

This simulation study is based on dichotomous item response data for an arithmetic test measuring 4 arithmetic operations (i.e. addition, subtraction, multiplication and division). Four subscores will be reported — one for each arithmetic operation. Two scenario-based testlets are built in this test, each with 10 items. Each testlet is constructed around a graph/plot/table on a given data scenario which requires examinees to conduct arithmetic operation(s) based on available information in the graph/plot/table. For example, a histogram depicting the frequency of students in a school for each ethnicity can be accompanied by questions like "What is the total number of students in this school?" (addition), "what is the proportion of White students in the school?" (division) and "what is the proportion of White and Asian students in the school?" (addition and division), etc. In addition to the two testlets, a set of another 10 items require information from both testlets, which is referred to *paired-testlets* in this study. In total, the test contains 30 items. It is assumed that all general abilities are independent from testlet effects, and testlet effects for the two testlets are correlated.

Manipulated Factors

To evaluate the proposed model across various study conditions, this simulation study manipulates 3 factors—(a) the magnitude of the testlet effect represented by the standard deviation (SD) of testlet effects (0.5, 1), (b) the correlation between the testlet effects of the dual testlets (0.2, 0.5, 0.8), and (c) the percentage of double-coded items in the test (20%, 40%, 60%). Table 2 summarizes the manipulated factors and their manipulated levels.

Table 2.
Manipulated Factors

Manipulated Factors	Level 1	Level 2	Level 3
SD of Testlet Effects	0.5	1	N/A
Correlation between Testlet Effects of the Dual Testlets	0.2	0.5	0.8
Percentage of Double-Coded Items	20%	40%	60%

The choice of the manipulated factors and the levels of manipulation are justified as follows.

Magnitude of testlet effects. The SD of the testlet effect parameters is an indicator of the magnitude of testlet effects in the current study. In Bradlow et al. (1999), they have altered the variance of the testlet effects, σ_γ^2, at 0.5, 1, 2 to achieve the ratio of the testlet effect variance to the sum of item difficulty variance and the testlet variance (i.e. $\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\sigma_b^2}$) of $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$. Wang and Wilson (2007) explored the performance of the Rasch testlet model at four evenly spaced testlet variances from 0 to 1(i.e. 0.25, 0.5, 0.75 and 1). Jiao et al. (2005) and Xie (2014) have compared model recovery with the SD of the testlet effects at 0, 0.5, 1 and 1.5. These studies have found a consistent pattern that as the variation of the testlet effects increases, the model parameter recovery is less accurate. In the investigation of the 2PL-DTM, Jiao et al. (2017) fixed the SD of testlet effect at 0.5.

Bradlow et al. (1999) found that the variance of testlet effect in a released SAT test is 0.11. Wang and Wilson (2005) conducted a real data analysis with the 2001 English test of the Basic Competence Tests for Journal High School Students in Taiwan and obtained testlet effect variance ranging from 0.007 to 2.09. According to previous findings, the SD of testlet effects in the current study are manipulated at 0.5 and 1, to represent small to moderate variation in testlet effects which are the

magnitude of the testlet effects often observed in real data analyses. The SD of testlet effects is the same for both testlets in this simulation study.

Correlation between testlet effects of the two testlets. As the paired testlets are based on information of the two testlets, it is reasonable to assume that the testlet effects for the first two testlets are correlated. Such correlation is manipulated at 0.2, 0.5 and 0.8 in this study to represent weak, moderate and strong correlation between the two paired testlets. The levels of manipulation are adopted from Jiao et al. (2017) for investigating the 2PL-DTM, since the proposed model also models paired testlets. Jiao et al. (2017) found a weak relation between the accuracy of parameter estimation and the correlation between paired testlet effects. However, since they only presented results from one replication for each study condition, the parameter recovery was judged by only bias, absolute bias, 95% credible interval capture rate. The stability of item and person parameter estimates were not assessed. The current study investigates the impact of the correlation between paired testlet effects on the model parameter recovery in terms of both estimation accuracy and the stability of the estimates.

Percentage of double-coded items. The percentage of double-coded items varies at 20%, 40% and 60% across study conditions. The levels of manipulation are to examine the capability of the proposed model in modeling item responses for tests with small, medium and large numbers of double-coded items. As the test involves testlets, double-coded items and their components, the percentage of double-coded items is selected to match with a specific structure of the test. The test structures are selected to ensure (a) a balanced assessment of all four arithmetic operations, (b) both double-coded items and their corresponding components are nested within the same

cluster of items (i.e. testlet or paired testlet), and (c) each testlet contains the same number of double-coded and single-coded items. Three simple test blueprints are developed to indicate the targeted arithmetic operation(s) for each item for tests consisted of 20%, 40% and 60% of double-coded items, respectively. The three simple test blueprints are presented in Table 3, jointly.

Table 3.

Test Structures for Tests with 20%, 40% and 60% of Double-Coded Item

Item	Testlet	20% Double-Coded Item	40% Double-Coded Item	60% Double-Coded Item
		Arithmetic Operation(s)	Arithmetic Operation(s)	Arithmetic Operation(s)
1	1	A (1)	A (1)	A (1)
2	1	A (2)	A (2)	S (2)
3	1	S (3)	S (3)	M (3)
4	1	S	S (4)	D (4)
5	1	M (5)	M (5)	A (1) & S (2)
6	1	M	M (6)	A (1) & M (3)
7	1	D	A (1) & S (3)	A (1) & D (4)
8	1	D	A (2) & S (4)	S (2) & M (3)
9	1	A (1) & S (3)	A (1) & M (5)	S (2) & D (4)
10	1	A (2) & M (5)	A (2) & M (6)	M (3) & D (4)
11	2	A (11)	A (11)	A (11)
12	2	A	A (12)	S (12)
13	2	S (13)	S (13)	M (13)
14	2	S	S (14)	D (14)
15	2	M (15)	D (15)	A (11) & S (12)
16	2	M	D (16)	A (11) & M (13)
17	2	D (17)	A (11) & D (15)	A (11) & D (14)
18	2	D	A (12) & D (16)	S (12) & M (13)
19	2	A (11) & D (17)	S (13) & D (15)	S (12) & D (14)
20	2	S (13) & M (15)	S (14) & D (16)	M (13) & D (14)
21	1 & 2	A	S (21)	A (21)
22	1 & 2	A	S (22)	S (22)
23	1 & 2	S (23)	M (23)	M (23)
24	1 & 2	S	M (24)	D (24)
25	1 & 2	M (25)	D (25)	A (21) & S (22)
26	1 & 2	M	D (26)	A (21) & M (23)
27	1 & 2	D (27)	S (21) & M (23)	A (21) & D (24)
28	1 & 2	D (28)	S (22) & M (24)	S (22) & M (23)
29	1 & 2	S (23) & D (27)	M (23) & D (25)	S (22) & D (24)
30	1 & 2	M (25) & D (28)	M (24) & D (26)	M (23) & D (24)

Note:1. A stands for addition, S stands for subtraction; M stands for multiplication, D stands for division.

2. For double-coded items, the number in the parenthesis for each targeted arithmetic operation indicates the source of the component item difficulty (i.e. the position of the component items). For component items, the number in the parenthesis indicates its position in the test. Standalone items have no parenthesis.

In the test with 20% of double-coded items, each testlet contains two single-coded items for each arithmetic operation. The test contains all possible combinations of double-coded items for the four operations, where each combination was tested with one item. In the test with 40% of double-coded items, each of all possible combinations is assessed by two items. The single-coded items are selected based on the double-coded items assigned to that testlet. In a test with 60% of double-coded items, each testlet has four single-coded items one for each arithmetic operation and each of all possible combinations is assessed by 6 double-coded items. By fully crossing levels of all manipulated factors, this simulation study contains 18 study conditions in total. Table 4 presents all 18 study conditions.

Table 4.
Simulation Conditions

Condition	SD of Testlet Effect	Correlation between Testlet Effects of Two Testlets	Percentage of Double-Coded Items
1	0.5	0.2	20%
2	0.5	0.2	40%
3	0.5	0.2	60%
4	0.5	0.5	20%
5	0.5	0.5	40%
6	0.5	0.5	60%
7	0.5	0.8	20%
8	0.5	0.8	40%
9	0.5	0.8	60%
10	1	0.2	20%
11	1	0.2	40%
12	1	0.2	60%
13	1	0.5	20%
14	1	0.5	40%
15	1	0.5	60%
16	1	0.8	20%
17	1	0.8	40%
18	1	0.8	60%

Fixed Factors

To conduct fair comparisons across the 18 conditions in Table 4, other factors are fixed in the simulation study. The study is based on a test that contains three clusters of items, each with 10 items. As indicated in Table 3, the first 10 items belong to the first testlet, items 11 to 20 belong to the second testlet, and the last 10 items are based on information from both the first and the second testlet.

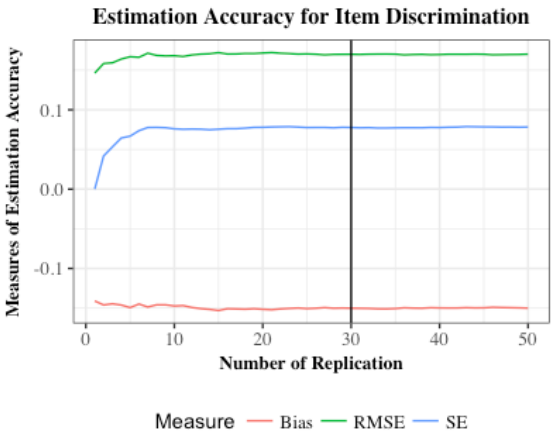
For all items, item discrimination is generated from a lognormal distribution with a mean of 0 and standard deviation of 0.5; item difficulties for single-coded items (i.e. stand-alone or component items) are generated from a normal distribution with a mean of 0 and standard deviation of 1. For the double-coded items, the component weights are generated from a uniform distribution with a minimum of 0 and a maximum of 1 to ensure that generated weights are between 0 and 1. Following Butter et al. (1998), the intercept parameter in the composite of the item difficulty is set at 0.5.

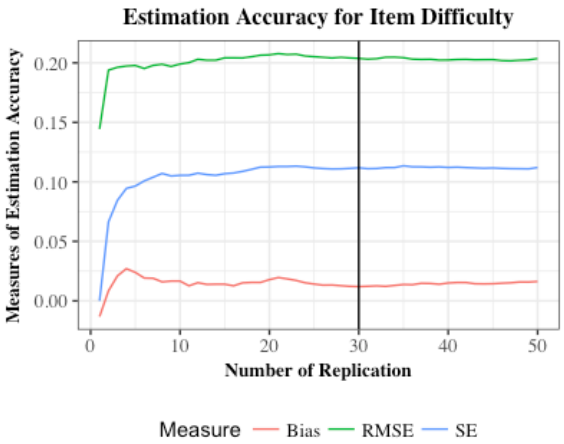
The person ability parameters and testlet effects are generated from a multivariate normal distribution with a mean vector of 0s, a variance of 1 for person ability, variances for testlets to the manipulated value, and the covariance depending on the manipulated testlet variance and the correlation between the paired testlets. The person abilities are independent of the testlet effects. Item responses are generated by comparing the calculated probability of obtaining a correct answer and a randomly generated value from a uniform distribution with a minimum of 0 and a maximum of 1. If the calculated probability is larger, then the simulated item response is 1. Otherwise, the item response is 0.

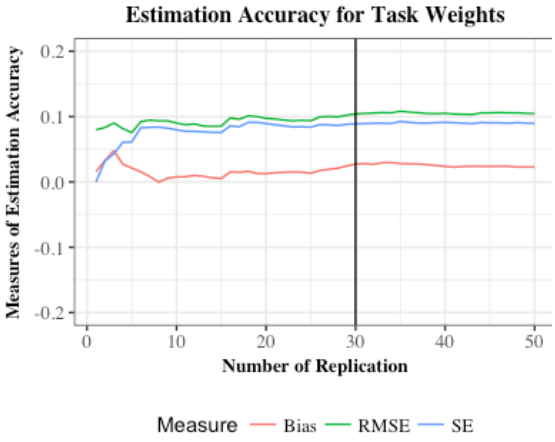
This study fixes the sample size to 1000. Normally, it is recommended that a sample size over 500 is appropriate for the Rasch model (Hambleton & Jones, 1993) and 1000 to stabilize item parameter estimates in a 2PL model. Simulation studies evaluating testlet models based on a 2PL IRT framework normally use sample sizes of 1000 or more (e.g. Bradlow et al., 1999; Jiao et al., 2005). Jiao et al. (2017) fixed the sample size at 2000 and have obtained reasonably accurate estimates for the proposed 2PL-DTM based on a state reading test with paired stimuli. From the perspective of the MIRID, although Butter et al. (1998) suggested that the parameter recovery accuracy was much improved for the dataset that contains 3000 examinees compared to that with 300 examinees, Wang and Jin (2010a) have successfully recovered the model parameters in a 2PL multilevel random weights MIRID with response data from 1000 examinees on 10 item families, each with 3 component items and 1 composite item.

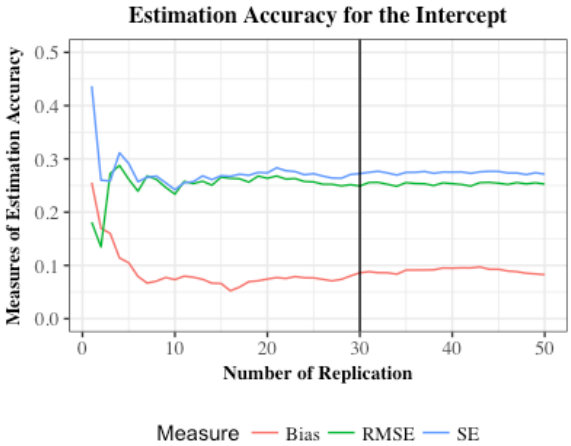
A pilot study was conducted to compare the average bias of each type of model parameters (i.e. item discrimination, item difficulty, task weights, intercept and overall ability parameters) with a sample size of 1000 and 2000 across other study conditions with 1 dataset from each condition. A t-test has found no significant difference in terms of the average bias for model parameters estimated with datasets containing 1000 examinees and that with 2000 examinees ($t_{abias} = -0.983, \mathrm{p_{abias}} = 0.326; t_{bbias} = -0.782, p_{bbias} = 0.434; t_{\sigma bias} = -0.629,\ p_{\sigma bias} = 0.531; t_{\tau bias} = 0.631,\ p_{\tau bias} = 0.532; and\ t_{\theta bias} = 0.729,\ p_{\theta bias} = 0.466$). Therefore, the sample size is fixed at 1000 examinees in this study.

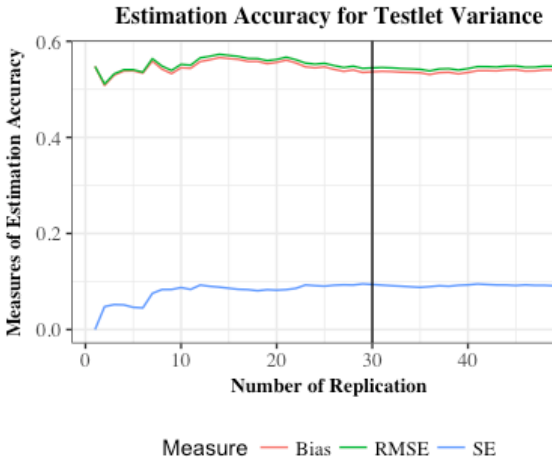
Each simulation condition is replicated for 30 times. Harwell, Stone, Hsu and Kirisci (1996) have investigated the number of replications needed in a Monte Carlo study with IRT models and recommended a minimum of 25 replications to maintain stable and small standard error for detecting an effect with a small effect size (i.e. $\eta^2 = 0.02$) (p. 111). Xie (2014) has shown that 30 replications are sufficient to obtain stable SEs for item difficulty estimates via a post hoc analysis using a multilevel cross-classified testlet model. A similar analysis was conducted to examine if 30 replications are sufficient to obtain stable estimates with the proposed model. In the pilot study, the proposed model and other competing models were fitted to 50 replicated datasets in the condition where the SD of the testlet effect is 0.5, the correlation between dual-testlet effects is 0.8, and there is 20% of double-coded items. Figure 5 shows the changes in the average bias, SEs and RMSEs for each type of parameters have reached stability at the 30th replication for the proposed 2PL-DT-MIRID. Parameters in other competing models (i.e. 2PL-DTM, 2PL-TMIRID, 2PL-MIRID, 2PL model) have also achieved stable estimates with fewer than 30 replications based on the average bias, SEs and RMSEs.

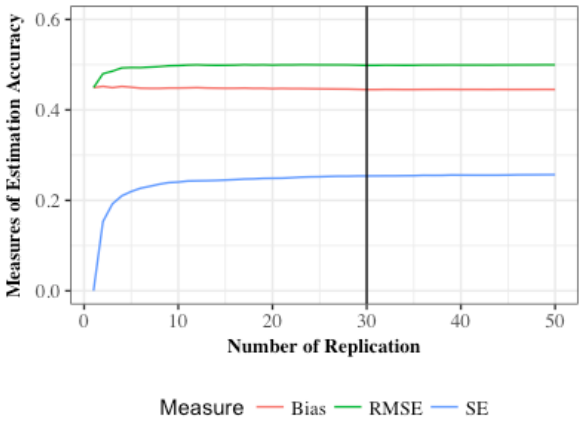

Estimation Accuracy for Item Discrimination
Measures of Estimation Accuracy
0.1
0.0
-0.1
0
10
20
30
40
50
Number of Replication
Measure
Bias
RMSE
SE


Estimation Accuracy for Item Difficulty
Measures of Estimation Accuracy
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
Number of Replication
Measure
Bias
RMSE
SE


Estimation Accuracy for Task Weights
Measures of Estimation Accuracy
0.2
0.1
0.0
-0.1
-0.2
0
10
20
30
40
50
Number of Replication
Measure
Bias
RMSE
SE


Estimation Accuracy for the Intercept
Measures of Estimation Accuracy
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
Number of Replication
Measure
Bias
RMSE
SE


Estimation Accuracy for Testlet Variance
Measures of Estimation Accuracy
0.6
0.4
0.2
0.0
0
10
20
30
40
Number of Replication
Measure
Bias
RMSE
SE


Estimation Accuracy for Overall Ability
Measures of Estimation Accuracy
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Number of Replication
Measure
Bias
RMSE
SE

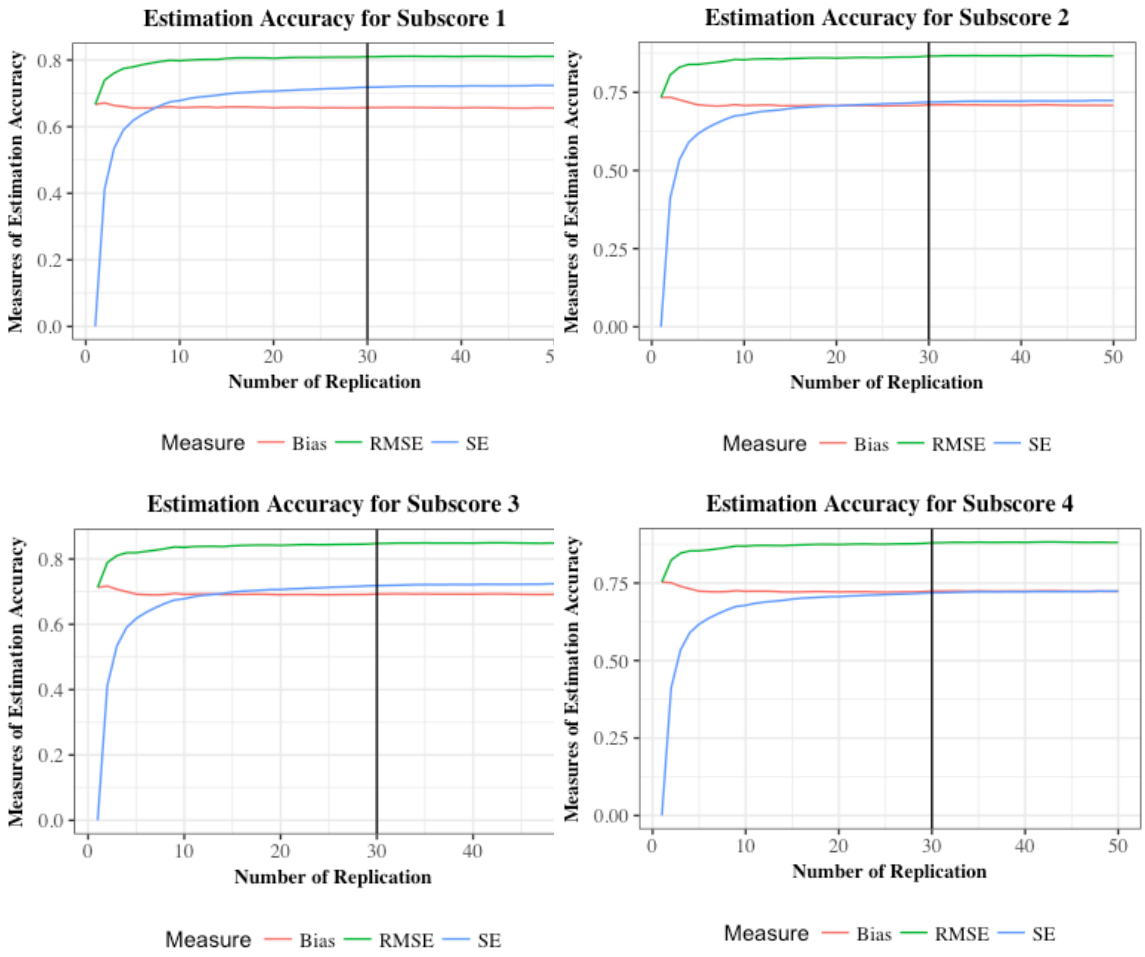


Figure 5. Bias, SE, RMSE for parameters as number of replication increases

The item and person ability parameters used to generate item responses are constant across replications within a condition. Doing so, the randomness in generating item responses within a study condition lies in comparing the randomly generated values from the uniform distribution with the calculated probability.

Data Generation

Data Generating Models

Test structures in Table 3 contain 6 types of items— (a) double-coded items in a paired-testlet, (b) component items in a paired-testlet, (c) double-coded items in a

single testlet, (d) component items in a single testlet, (e) stand-alone items (i.e. a single-coded item that does not serve as a component item for any double-coded item) in a paired-testlet and (f) stand-alone items in a single testlet.

For double-coded items in a paired testlet, the proposed model in Equation 28 should be used. For the k^{th} component of item i in paired testlets, the probability of obtaining a correct answer is modeled using the 2PL-DTM proposed by Jiao and Lissitz (2014) and Jiao et al. (2017). The item subscript i for such component item takes on the same value as in the double-coded/composite item that the component item contributes to. Another way to understand the subscript is that the double-coded item and its corresponding component items can be considered as an item family. The subscript i is an item family indicator. Equation 29 presents the Jiao and Lissitz (2014) model with subscripts adapted for the composite-component situation.

$$\mathrm{P}(\mathrm{X_{jik}}=1)=\frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)} * \frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_2(i)}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_2(i)}-\beta_{ik}\right)\right)}. \quad (29)$$

For a double-coded item i with K components in a single testlet— for example, testlet d_1, the probability person of j $(j=1\ldots J)$ obtaining the correct answer is modeled as follows:

$$\mathrm{P}(\mathrm{X_{ji0}}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}. \quad (30)$$

For a component k of item i in a single testlet, for example testlet d_1, the probability of obtaining a correct answer is modeled with a regular two-parameter testlet model (2PL-TM) that is presented as follows:

$$\mathrm{P}(\mathrm{X_{jik}}=1)=\frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}. \tag{31}$$

A stand-alone test item is treated as a special case of a composite item where there is only one component, and the component difficulty is weighted by 1. The reason of making the stand-alone as a special case of a composite item is for the convenience of looping in data generation and data estimation. The probability for answering a stand-alone item impacted by both testlets correctly can be represented as

$$\begin{aligned}\mathrm{P}(\mathrm{X_{ji0}}=1)=&\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sigma_1\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sigma_1\beta_{i1}\right)\right)}\\ &*\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\sigma_1\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\sigma_1\beta_{i1}\right)\right)},\end{aligned} \tag{32}$$

where $\sigma_1=1$.

Similarly, the probability of obtaining a correct answer for a stand-alone item nested in testlet d_1 is modeled as

$$\mathrm{P}(\mathrm{X_{ji0}}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{i1}\right)\right)}. \tag{33}$$

Table 5 presents true models (i.e. data simulating models) for different types of items in the designed test. Appendix A summarizes the data generating models for all items in a test with 20% double-coded items for demonstration. Based on models in Table 5, dichotomous item responses are generated using R.

Table 5.
Data Simulating Models

Item Type	Testlet Type	Model
Double-Coded	Paired-Testlet	Proposed 2PL-DT-MIRID (Equation 28)
Component	Paired-Testlet	2PL-DTM (Equation 29)
Double-Coded	Single	2PL-T-MIRID (Equation 30)
Component	Single	2PL-TM (Equation 31)
Standalone	Paired-Testlet	A special case of 2PL-DT-MIRID with only one component; the component difficulty is weighted by 1. (Equation 32)
Standalone	Single	A special case of 2PL-T-MIRID with only one component, the component difficulty is weighted by 1. (Equation 33)

Subscores are not model parameters in data generating models. The subscores (i.e. subdomain ability) are computed by using the generated item discrimination parameters, the item difficulty parameters for the content domain, the weights for the content domain, the testlet parameters and the item responses. For example, the calculation of subscore of addition uses item responses to items that assess addition and true item parameters for those items. For a double-coded item that assesses addition and another arithmetic operation, only the weight and the component item difficulty for addition is used. The empirical true subscores are the averages of the computed subscores across replications within each condition.

Model

Model Comparison

To evaluate the consequences of ignoring the double-coded item and/or the effect of the paired testlets, this simulation study also compares model parameters and

sub-content domain ability estimates for scenarios where underspecified models are used to model item responses for complex innovative item types. Specifically, for each dataset in each condition, the true models are first used to estimate model parameters and sub-content domain scores, then the underspecified models are applied to estimate model parameters and subscores. When estimating subscores, item parameters used are those obtained by using different competing models in model parameter estimation. Estimated subscores are compared across the following competing models.

Comparison model 1: true models. Subscores are estimated using the data generating models as presented in Table 5.

Table 6.
Models Used for Ignoring Paired-Testlet Effect

Item Type	Testlet Type	Model
Double-Coded	Paired-Testlet	2PL-T-MIRID (Equation 30)
Component	Paired-Testlet	2PL-TM (Equation 31)
Double-Coded	Single	2PL-T-MIRID (Equation 30)
Component	Single	2PL-TM (Equation 31)
Standalone	Paired-Testlet	A special case of 2PL-T-MIRID with only one component, the component difficulty is weighted by 1. (Equation 33)
Standalone	Single	A special case of 2PL-T-MIRID with only one component, the component difficulty is weighted by 1. (Equation 33)

Comparison model 2: ignoring the effect of paired testlet. Ignoring the effect of the paired testlets, the double-coded items embedded in paired testlet is fitted with a 2PL-T-MIRID, the component items in the paired testlet is modeled by a 2PL-TM, and the stand-alone item in the paired testlet is modeled by the special case of

2PL-T-MIRID. In other words, items belonging to the paired testlet are now considered nested within a third testlet. Table 6 presents the models used to estimate domain scores in comparison model set 2 where the effect of paired testlet is ignored.

Comparison model 3: ignoring the testlet effect. Testlet effects are completely ignored in this scenario. Regardless of testlet membership, all double-coded items are modeled with 2PL-MIRID and the stand-alone items are fitted with the 2PL model.

Item responses to component items are modeled with the 2PL adapted for the composite-component situation (See Equation 34). Table 7 presents models used in comparison model set 3.

$$P(X_{jik} = 1|\theta_i) = \frac{\exp(a_{ik}(\theta_i - \beta_{ik}))}{1 + \exp\,(a_{ik}(\theta_i - \beta_{ik}))}. \quad (34)$$

Table 7.
Model Used in Ignoring Testlet Effects

Item Type	Testlet Type	Model
Double-Coded	Paired-Testlet	2PL-MIRID (Equation 27)
Component	Paired-Testlet	2PL in MIRID (Equation 34)
Double-Coded	Single	2PL-MIRID (Equation 27)
Component	Single	2PL in MIRID (Equation 34)
Standalone	Paired-Testlet	2PL (Equation 2)
Standalone	Single	2PL (Equation 2)

Comparison model 4: ignoring double-coded items. The double-coded items are treated as single-coded items in this scenario. In other words, all items in the test are single-coded. Therefore, items in the paired testlet are modeled using Jiao

and Lissitz (2014), items in the single testlet are modeled with a simple 2PL-testlet model (Bradlow et al., 1999). Table 8 presents models used for scenario 4.

Table 8.
Model Used in Ignoring Double-Coded Item

Item Type	Testlet Type	Model
Double-Coded	Paired-Testlet	2PL-DTM (Equation 15)
Component	Paired-Testlet	2PL-DTM (Equation 15)
Double-Coded	Single	2PL-TM (Equation 10)
Component	Single	2PL-TM (Equation 10)
Standalone	Paired-Testlet	2PL-DTM (Equation 15)
Standalone	Single	2PL-TM (Equation 10)

Comparison Model 5: ignoring double-coded items and testlet effects. The double-coded items are treated as single-coded items, and the testlet effects are ignored in this model set. In other words, all items in the test are fitted with a 2PL model. Table 9 presents models used for comparison model set 5.

Table 9.
Model Used in Ignoring Double-Coded Item Structure and Testlet Effect

Item Type	Testlet Type	Model
Double-Coded	Paired-Testlet	2PL (Equation 2)
Component	Paired-Testlet	2PL (Equation 2)
Double-Coded	Single	2PL (Equation 2)
Component	Single	2PL (Equation 2)
Standalone	Paired-Testlet	2PL (Equation 2)
Standalone	Single	2PL (Equation 2)

In comparison model sets 1 to 5, the subscores are estimated following a two-step procedure. First, the item parameters are estimated for all items in a single

calibration. The overall ability parameters are estimated. Then, the four sub-content domain scores are estimated for each examinee using the item parameters calibrated from the first step, the estimated testlet effects and based on item responses for the subset of items targeting on the sub-content domain. When including the double-coded items, only the item difficulties for that specific sub-content domain are used for estimating subscores.

Comparison model 6: IRT-based NCS. As described previously, the proposed model can be used to report subscores using IRT-based NCS. The IRT-based NCS subscores are estimated with the following steps.

(a) Calibrate item parameters using the data generating model.

(b) Formulate the TCC for each subtest consisting of items testing on a specific arithmetic operation. The testlet effects are integrated out in the computation of the test characteristic function. For double-coded items, only difficulties for that arithmetic operation are used in formulating the TCC.

(c) Calculate the sum subscores for each student on each sub-content domain.

(d) Solve the TCC for sub-content domain ability for each examinee.

Model Identification

The interaction between person ability and item difficulty in an IRT model leads to scale indeterminacy if no constraint is applied. A common approach to set the scale for $(\theta_j - b_i)$ is to constrain the mean of person abilities or the item difficulties to be 0. The current study involves decomposition of the item difficulty parameters for the double-coded items. Specifically, the item difficulty for a double-coded item is a weighted sum of the component difficulties. Therefore, constraining the mean of

item difficulties to be 0 in the current study is not straightforward. Hence, the mean of the person abilities is constrained to be 0. For model identification, the person ability parameters are assumed from a standard normal distribution in Bayesian estimation.

Model Parameter Estimation

The current study uses the Bayesian estimation method to estimate model parameters for the proposed model and other models used in the comparison scenarios. Specifically, the MCMC algorithm is applied to parameter estimation. Due to the popularity of the Bayesian estimation method, many software programs are developed for models with various specifications. Commonly used Bayesian software include but are not limited to Stan (Carpenter et al., 2016), JAGS (Plummers, 2017), WinBUGS (Spiegelhalter et al., 2003) and OpenBUGS (Spiegelhalter, Thomas, Best & Lunn, 2004). In this study, JAGS is used for model parameter estimation.

In using the Bayesian estimation method, priors must be specified for each estimated parameter. In this study, the prior for the person abilities is a standard normal distribution with a mean of 0 and standard deviation of 1 (i.e., $\theta_j \sim N(0,1)$) for scale identification; the prior for the component item difficulties is a normal distribution with a mean of 0 and variance of 2 (i.e. $\beta_{ik} \sim N(0,2)$), as such flat prior will have less influence on the results and allow data to be weighed more in estimating model parameters; the prior for the component weights is specified as a uniform distribution with a minimum of 0 and a maximum of 1(i.e. $\sigma_k \sim (0,1)$); the prior for the intercept is a standard normal distribution (i.e. $\tau \sim N(0,1)$); and the item discrimination parameters have a prior of lognormal distribution with a mean of 0 and a variance of 0.5 (i.e. $a_{ik} \sim Lognormal\ (0,\ 0.5)$), since the reasonable values of item

discrimination parameters are within the range of (0, 2) and larger values of item discriminations within this range are desired. A multivariate normal distribution is assumed for the testlet effects. The mean vector contains 0s and the variance and covariance matrix follows an inverse Wishart distribution, with 1s set as the priors for the variances and 0s for the covariances (i.e. $\gamma_{jd_t(i)} \sim MVN\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \sim \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\Sigma \sim \mathrm{W}^{-1}\ (\boldsymbol{\psi}, \boldsymbol{\nu})$ in which $\nu = 2$ and $\boldsymbol{\psi} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. As the conjugate prior for the multivariate normal distribution, the inverse Wishart distribution is used as a hyper-prior to integrate out the unknown covariance matrix in the prior so as to estimate the variance-covariance matrix for the multivariate normal distribution of testlet effects as in the posterior distribution.

In estimating model parameters using the MCMC algorithm, unknown parameters are drawn from the posterior distributions via Gibbs sampler. To sample parameters for the proposed model, suppose $\boldsymbol{\theta}$ is a vector of abilities for all *J* examinees taking the test, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$; let $\boldsymbol{\beta}$ be a vector of component difficulties, where $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \ldots, \beta_{IK})$, let $\boldsymbol{\sigma}$ be a vector of component weights, in this case, we have 4 components in the test— $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$; $\boldsymbol{\gamma}$ is a vector of testlet effects where $\boldsymbol{\gamma} = (\gamma_{1d_1(i)}, \gamma_{1d_2(i)}, \ldots, \gamma_{Jd_2(i)}$; the variance and covariance matrix of the testlet effect is denoted as $\boldsymbol{\sigma_\gamma^2}$; $\boldsymbol{a}$ is a vector of item discrimination parameters where $\boldsymbol{a} = (a_{11}, a_{12}, \ldots, a_{IK})$, and τ is the intercept in the item difficulty composite. Hence, $\boldsymbol{\omega}$ (i.e. $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\sigma_\gamma^2}, \tau)$) determines item responses to items in the test. Through Gibbs sampler, $\boldsymbol{\omega^t}$ is updated to $\boldsymbol{\omega^{t+1}}$. After the change of the parameters is minimal and stable, and chains with different starting values mixing

well, convergence has been achieved. The estimates are obtained by averaging the values of $\boldsymbol{\omega}$ after convergence has been reached for each estimated parameter. In this study, convergence was checked by observing the diagnostic plots based on results from JAGS, such as traceplot, quantile plots, etc. In addition, the Gelman-Rubin convergence statistic $\hat{R}$, as modified by Brooks and Gelman (1998), was also calculated. Values of $\hat{R}$ less than 1.1 indicate convergence (Brooks & Gelman, 1998).

A pilot study was conducted using datasets generated for a test with 20% double-coded items, testlet effect standard deviation of 0.5, a correlation of 0.8 between testlet effects for the first and second testlets. The MCMC estimation method in JAGS ran two chains with 20,000 iterations for burn-in for the true models. Since the R package R2JAGS has been used to implement MCMC estimation in JAGS, R functions have been developed to estimate parameters for the proposed model and other competing models.

Analysis

Parameter Recovery Accuracy

The parameter recovery accuracy is evaluated for item parameters (i.e. item discrimination, component difficulties, component weights, composite intercept) and person parameters (i.e. testlet effect and person ability parameters) by comparing the estimates to the true model parameters, if the parameters appear in the proposed model or in other comparison models. Bias, SE and RMSE are used as indicators of model parameter recovery accuracy. These indices were selected because they

address different perspectives of parameter estimation accuracy. Used together, they provide comprehensive assessment of model parameter recovery.

Bias. Bias is an index for systematic error. The equation for calculating bias for parameter ξ is presented in Equation 35. The difference between the parameter estimate and the value of the true parameter are averaged across replications, *N*. In other words, bias is an average of how much the parameter estimate deviates from the true value of the parameter.

$$\text{Bias}(\hat{\xi}) = \frac{\sum_{\text{r=1}}^{\text{N}}(\hat{\xi}_r - \xi)}{\text{N}}. \tag{35}$$

Standard error (SE). The SE is an index assessing random error in estimation. The equation for SE is presented in Equation 36. It is a measure of how much the parameter estimate deviates from the average of parameter estimates across all replications. The estimated parameter in replication l ($l = 1,2, \dots, N$) or r ($r = 1,2, \dots, N$) is denoted as $\hat{\xi}_l$ or $\hat{\xi}_r$.

$$\text{SE}(\xi) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\xi}_r - \frac{\sum_{l=1}^{N}\hat{\xi}_l}{N})^2}. \tag{36}$$

Root mean squared error (RMSE). The RMSE is a measure of total error in parameter estimation. The RMSE is defined in Equation 37. The calculation of RMSE captures both the bias and the variability of the parameters.

$$RMSE(\xi) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\xi}_r - \xi)^2}. \tag{37}$$

The bias, SEs and RMSEs are averaged across ability parameters, testlet variance, item difficulty parameters, item discrimination parameters and weights respectively when investigating the impact of manipulated factors on parameter

recovery. For example, in a study condition where there are 1000 examinees, the bias, SEs and RMSEs for the 1000 ability estimates are averaged and reported.

To evaluate the impact of the manipulated factors on the bias, SE and RMSE, this study conducts analysis of variance (ANOVA) for the bias, SEs and RMSEs obtained based on Equation 35, 36 and 37, respectively. The purpose of ANOVA is to compare parameter recovery across study conditions and across comparison model sets for statistical inferences. In ANOVA, the alpha level is set at 0.05 for statistically significant difference. For statistically significant factors, the partial eta square, denoted as η_p^2, is reported as effect size measure. The partial eta square is defined in Equation 38, where the sum of squares of the investigated factor is divided by the total variance of the dependent variable after the effects of other independent variables and interactions have been partialled out (Cohen, 1965). There is no rule of thumb for what is defined as small, medium and large effects when using eta squared (Richardson, 2011). The recommended values in Cohen (1969) is selected as criteria in this study— $\eta_p^2 = 0.01$ for small effect, $\eta_p^2 = 0.06$ for medium effect and $\eta_p^2 = 0.14$ for large effect.

$$\eta_p^2 = \frac{SS(A)}{SS(A) + SS(within)}. \tag{38}$$

The ANOVA is not conducted for task weights (i.e. σ_k $(k = 1,2,3,4)$), intercept (i.e. τ), testlet variances (i.e. VAR_{γ_1} and VAR_{γ_2}) and the correlation between the dual testlets (i.e. $\rho_{\gamma_1\gamma_2}$), because the sample size is insufficient to assess assumptions or to have enough power for the analysis. The ANOVA is only conducted for the bias, SE, and RMSE of the item discrimination parameters (i.e. $a_i, i = 1,2,\dots,30$), the item difficulty parameters (i.e. $b_i, i = 1,2,\dots,30$), the overall

ability parameters (i.e. $\theta_j, j = 1,2,\ldots,1000$) and the subcores (i.e. $\theta_{jA}, \theta_{jS}, \theta_{jM}\ and\ \theta_{jD}, j = 1,2,\ldots,1000$). The manipulated factors (i.e. testlet effect SD, correlation between testlet effects of the paired testlets and the percentage of double-coded items) are treated as fixed effects in the ANOVA. Since all competing models are fitted to each of the generated response dataset, the models are treated as repeated measures in the ANOVA design.

Whether the sphericity assumption, normality assumption and homogeneity assumption are met with these data are assessed. When the sphericity assumption is violated, a Huynh-Feldt adjustment is applied to the degrees of freedom to adjust for inflated Type I error in the F test. Although the normality assumption is checked for ANOVA, no adjustment is made. This is because studies have found that the impact of non-normality on the Type I error rate is minimal in an F-test (Glass, Peckham & Sanders, 1972). The plausibility of the homogeneity assumption is checked for error measures of all parameters using the Levene's test. For a parameter that has the same number of parameters in each study condition (i.e. 30 item discrimination parameters in each condition, 1000 overall ability parameters in each study condition and 4*1000 subscores in each condition), ANOVA was conducted even if the homogeneity assumption is not met. This is because the impact of the violation of homogeneity assumption is minimal when the sample sizes are equal (Hinkle, Wiersma, & Jurs, 1998). However, the situation for item difficulty is different. As the estimates of composite item difficulty contain estimation errors from task weights and the intercept in addition to those from the item component parameters, the error measures (i.e. bias, SE and RMSE) for item difficulty are only summarized for single-coded

items. Since the percentage of double-coded items is manipulated and the total number of items remains the same, the number of bias, SEs and RMSEs for item difficulty parameters are not identical across study conditions, i.e. the equal sample size across cell is not satisfied. Hence, ANOVA results of item difficulty are only reported for the error measures that meet the homogeneity assumption.

Significant effects with at least a small effect size are summarized and reported. Pairwise comparison is planned to be conducted for a main effect when the following three conditions are met simultaneously: (a) the main effect is significant, (b) the main effect has at least a small effect, and (c) it does not have statistically significant interaction with other factors. This is because multiple comparisons generalize differences between levels of a main effect at the all-sample level (i.e. across all study conditions), but when the significant interaction effect is present, the impact of the main effect differs at different levels of the other effect(s). This study is not only interested in the effect of a factor across all study conditions; more importantly, it investigates how factors behave in different conditions. Therefore, multiple comparisons are not conducted for significant main effects when interaction effects are significant. The Dunn-Sidak procedure, which is more powerful than the Bonferroni procedure, is used to adjust for family-wise Type I error in the multiple comparison procedures for the within factor.

Score Reliability

The reliability was defined in the CTT framework by Lord and Novick (1968, p.61)— the reliability for test scores equals to one minus the ratio of error score variance to the total score variance.

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_X^2}. \quad (39)$$

In the IRT framework, reliability is not defined as a one-number index for a test score. Instead, reliability is calculated conditioning on person ability using IRT (Sireci, Thissen, & Wainer, 1991). In the current study, reliability is conditioned on the testlet effects for the two testlets as well as the person ability. When the true models are used, the marginal reliability, defined in Equation 40, is calculated using the expectation value of the measurement error variance obtained by integrating out the person ability and testlet effects.

$$\rho_{xx'} = 1 - \frac{\bar{\sigma}_e^2}{\sigma_X^2}, \quad (40)$$

where $\bar{\sigma}_e^2$ is the expectation value of $\sigma_e^2 = \left(\frac{1}{I}\right)^2$ (i.e. $\bar{\sigma}_e^2 = \int \left(\frac{1}{I}\right)^2 g(\theta_j, \gamma_{jd_1}, \gamma_{jd_2}) d\theta_j d\gamma_{jd_1} d\gamma_{jd_2}$).

Based on Equation 40, the marginal reliability for the test score and the four sub-content domain scores are calculated and compared across study conditions and across scenarios. Item information was derived for the proposed model based on the definition of test information in Equation 7. The derivation of test information is presented in Appendix B. R functions are developed to calculate the test score reliability for overall ability and subscores. Subscore reliabilities are calculated based on the subscores estimated in JAGS using estimated model parameters.

Model Selection

To select the best fitting model for the generated data, three model fit indices are calculated, including deviance information criterion (DIC; Spiegelhalter, Best,

Carlin, & van der Linden, 2002), Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwartz, 1978).

Deviance information criterion. The DIC is an index for assessing model fit with Bayesian posterior estimates. It measures the model adequacy and penalizes additional complex terms added in the model (Spiegelhalter, 2002). Equation 41 represents the mathematical formulation of DIC.

$$DIC = \overline{D(\xi)} + p_D = D(\bar{\xi}) + 2p_D, \tag{41}$$

where $\overline{D(\xi)}$ is the posterior mean deviance; $p_D = \overline{D(\xi)} - D(\bar{\xi})$, and $D(\bar{\xi})$ is the deviance at the posterior estimates of the parameter. Since larger values of DIC indicate worse model fit, p_D is the number of effective parameters in the model. The DIC index is requested directly from the JAGS in estimating model parameters.

Akaike information criterion. The AIC is calculated as in Equation 42.

$$AIC = -2lnL + 2P, \tag{42}$$

where lnL is the log likelihood and P is the number of parameters to be estimated. Smaller AIC value is desired in model comparison. The larger the log likelihood is, the smaller the AIC value is. Like DIC, AIC penalizes models with more parameters.

Bayesian information criterion. The BIC is another likelihood-based model fit index (see in Equation 43).

$$BIC = -2lnL + \ln(N) \cdot P. \tag{43}$$

In the calculation of BIC, the number of parameters, denoted as P, is weighted by the natural log of the number of observations in the data, denoted as $\ln(N)$. Therefore, BIC imposes a more severe penalty for complex models than AIC or DIC does.

All of these indices consist of two elements— the deviance of the model fit and the penalty term for model complexity. Among these three indices, BIC penalizes the more parameterized model the most.

For each replication within each condition, the three indices are calculated for each comparison model set. Comparison of the model fit is conducted using the proportion selecting the true model as the best fitting model with three model fit indices within each study condition.

Chapter 4: Results

The simulation study evaluates the proposed model in terms of (a) model parameter and subscore estimation accuracy, (b) score reliability and (c) model selection, in comparison with other under specified models described in Chapter 3. The comparison of models is carried out at different levels of the manipulated factors— (a) the testlet effect SD (0.5, 1), (b) the correlation between the testlet effects of the paired testlets (shorted as dual testlets correlation, 0.2, 0.5, 0.8) and (c) the percentage of double-coded items (20%, 40%, 60%). As described in Chapter 3, models that are used for items in a test could be different based on the item structure and testlet structure. For simplicity and clarity in the result summary, the name of the model that is used for a double-coded item requiring information from the paired testlets in each model set is used to identify the model set. Table 10 presents the model names used to represent each model set.

Table 10.
Model Sets and Their Abbreviated Names

Comparing Model Set	Name Used in Result Summary
True Model Set	DT-MIRID
Models ignoring dual-testlet structure	T-MIRID
Models ignoring all testlet effects	MIRID
Models ignoring double-coded item structure	DTM
Models ignoring testlet effects and the double-coded item	2PL
Number-correct scoring	NCS

Since parameters in the data generating model are not always in other underspecified models, parameter estimates are compared among models that contain the parameter being compared. Table 11 presents models being compared for different model parameters and subscores.

Table 11.
Comparing Models for Different Model Parameters and Subscores

Parameter	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
a_i	✓	✓	✓	✓	✓	
b_i	✓	✓	✓	✓	✓	
σ_k	✓	✓	✓			
τ	✓	✓	✓			
VAR_γ	✓	✓		✓		
$\rho_{\gamma_1\gamma_2}$	✓			✓		
θ_j	✓	✓	✓	✓	✓	✓
θ_{jA} (Subscore of Addition)	✓	✓	✓	✓	✓	✓
θ_{jS} (Subscore of Subtraction)	✓	✓	✓	✓	✓	✓
θ_{jM} (Subscore of Multiplication)	✓	✓	✓	✓	✓	✓
θ_{jD} (Subscore of Division)	✓	✓	✓	✓	✓	✓

All model parameters in each comparing model are converged in all study conditions. Parameter estimates in the DT-MIRID converged after the first 100,000 iterations. Model parameter estimates in the T-MIRID, the MIRID and the 2PL are converged after the first 5,000 iterations. The convergence for the model parameters in DTM has been achieved after the first 350,000 iterations. Samples before convergence are discarded as burn-in iterations. For all models, model parameter inferences are based on another 10,000 iterations after burn-in for each of the two chains. The chains of the DT-MIRID are thinned by 10, those of the DTM by 15, and chains of other models by 1.

The first section of this chapter presents the evaluation of parameter estimation accuracy. In the ANOVA, significant effects with at least small effect sizes on the bias, SE and RMSE are presented. Higher-order significant interactions are primarily explained as they are more meaningful for understanding how manipulated factors and model structure impact parameter estimates. As the lower-order

significant interactions overly generalize patterns found in significant higher-order interactions, they are not elaborated when higher-order interactions are statistically significant. However, this study does not explain significant four-way interactions. for two reasons—(a) since there are four factors in the ANOVA, a significant four-way interaction means that patterns are different in each study condition, and (b) the four-way interaction is too complicated to be interpreted in a meaningful way. Although not explained in detail, the four-way interaction is still included in the ANOVA design. The error variances in the F-test is based on the full ANOVA model. For parameters that are not assessed by ANOVA, key findings based on the marginal difference of factors are presented for each error measure. Bias, SEs and RMSEs for each model under each study condition are tabulated in Appendix C. The second section compares the score reliability yielded from different competing models across study conditions and provide possible explanation for the differences in the patterns observed for the reliabilities across study conditions. The third section presents results on model fit indices.

Parameter Estimation

Item Discrimination

Bias. The significant effects on bias of $\hat{a}_i$ with at least a small effect size are tabulated in Table 12. The three-way interaction among model, testlet effect SD and the dual testlet correlation is statistically significant with a small effect size ($\eta_p^2 = 0.037$). Figure 6 shows that the variability of mean bias of $\hat{a}_i$ at various levels of dual testlet correlation is larger when the testlet effect SD is 1 for all models. In addition,

when testlet effect variability is large, the magnitude of the bias for $\hat{a}_i$ is always the smallest when the dual testlets are less correlated, across all models. Whereas in the situation where the testlet effect SD is 0.5, the smallest bias of $\hat{a}_i$ is obtained when the dual testlets correlation is 0.5 for DT-MIRID and DTM and when the dual testlets correlation is 0.2 for T-MIRID, MIRID and 2PL. The three-way interaction effect among model, testlet effect SD and the percentage of double-coded items (see in Figure 7) is also significant with a small effect size ($\eta_p^2 = 0.011$). For DT-MIRID, T-MIRID, MIRID and 2PL, the magnitude of bias of $\hat{a}_i$ is the smallest when there are 60% of double-coded items in the test, at the lower level of testlet effect SD; whereas the bias of $\hat{a}_i$ is the smallest when 40% of double-coded items are in the test, at the higher level of testlet effect SD. For DTM, the bias for $\hat{a}_i$ is always the smallest when there are 20% of double-coded items regardless of the testlet effect SD. This is because the DTM ignores the double-coded item structure, so the impact on the bias on $\hat{a}_i$ estimated by DTM is the smallest when there are fewer double-coded items in the test.

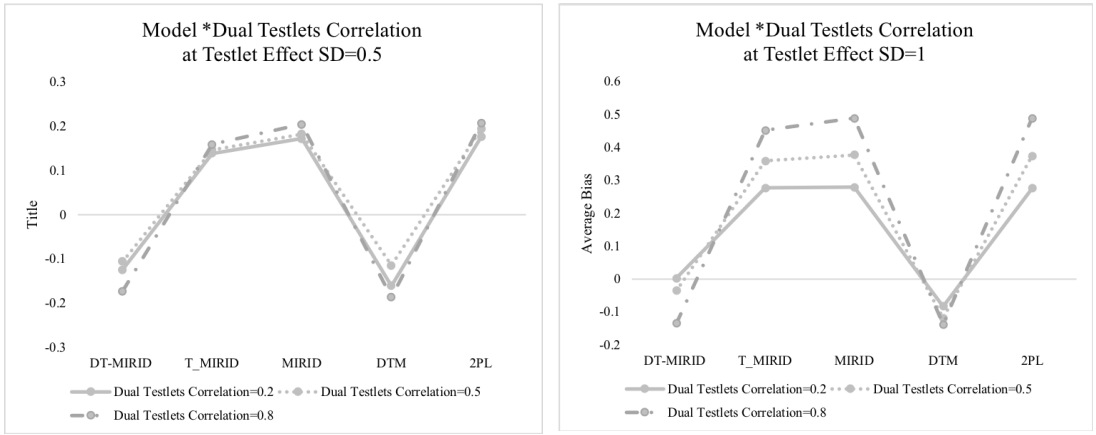


Figure 6. Significant three-way interaction on bias of $\hat{a}_i$—model* dual testlets correlation* testlet effect SD

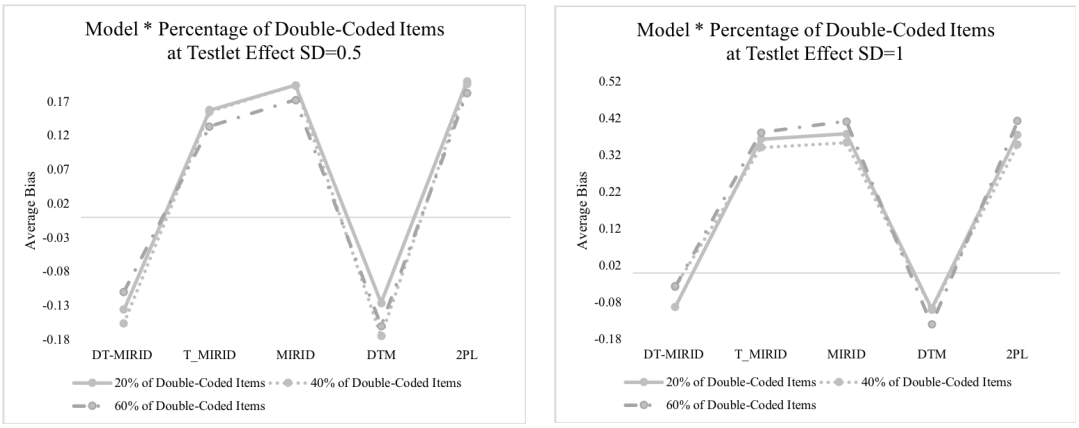


Figure 7. Significant three-way interaction on bias of $\hat{a}_i$—model* percentage of double-coded items* testlet effect SD

Table 12.
ANOVA Results of Significant Effects on the Bias of $\hat{a}_i$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	1255.617	<0.001	0.706
model * testlet.sd	41.412	<0.001	0.074
model * dbcorr	24.427	<0.001	0.086
model * testlet.sd * dbcorr	9.94	<0.001	0.037
model* testlet.sd*percent_dbcd	2.887	0.042	0.011
Between			
testlet.sd	132.477	<0.001	0.202
dbcorr	3.933	0.02	0.015
testlet.sd * dbcorr	3.627	0.027	0.014

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

In addition to the significant three-way interactions, three two-way interactions have significant effects on the bias of $\hat{a}_i$. The two-way interaction effects between model and testlet effect SD ($\eta_p^2 = 0.074$) and that between model and dual testlet correlation ($\eta_p^2 = 0.086$) are of medium effect size. The two-way interaction between testlet effect SD and dual testlet correlation is statistically significant with a small effect size ($\eta_p^2 = 0.014$). In terms of main effects, the main effects of model ($\eta_p^2 = 0.706$), testlet effect SD ($\eta_p^2 = 0.202$), and dual testlets correlation ($\eta_p^2 =$

0.015) are statistically significant with large, large, and small effect on bias of $\hat{a}_i$, respectively.

Because not all items in the test require information from the paired testlets, an investigation is conducted at the item level for the bias of $\hat{a}_i$. Such investigation reveals that $\hat{a}_i$ for items that require information from the paired testlets are positively biased in a much larger magnitude when the paired testlets are ignored, as compared to items nested within a single testlet. This pattern is consistent across study conditions. Figure 8 presents the bias of $\hat{a}_i$ in condition 10, where the testlet effect is 1, the dual testlets correlation is 0.2 and there are 20% of double-coded items. (Study condition 10 is chosen for demonstration is because this condition has produced relatively small bias and RMSEs of $\hat{a}_i$ when estimated with the proposed model). In Figure 8, the orange, grey and light blue bars respectively representing the bias of $\hat{a}_i$ yielded from the T-MIRID, the MIRID and the 2PL are much longer than the dark blue and yellow bars that represent the bias from DT-MIRID and DTM, for items 21-30. Moreover, the bias of $\hat{a}_i$ yielded from models ignoring dual testlets are always positive for items 21-30, whereas the $\hat{a}_i$ based on DT-MIRID and DTM for those items are negatively biased.

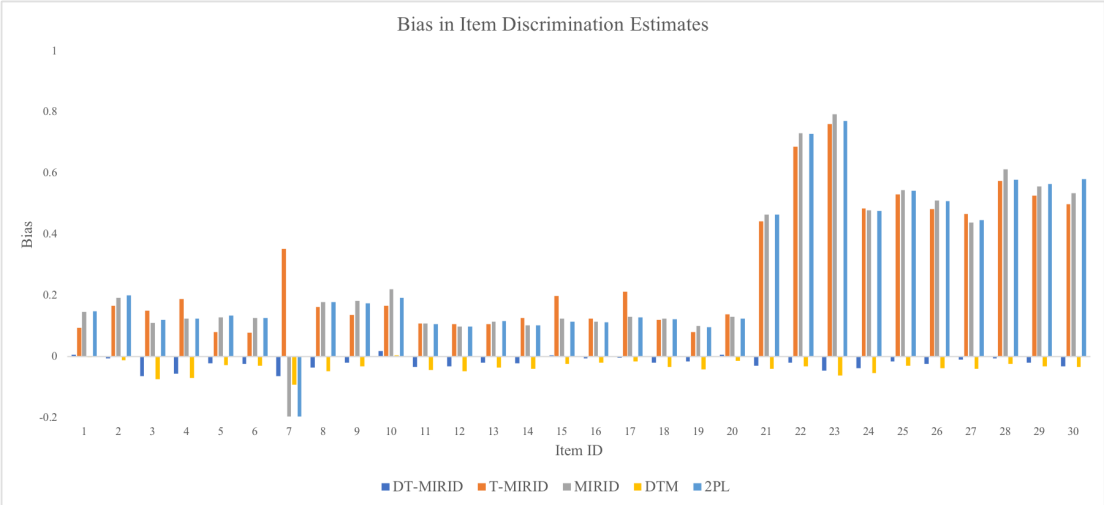


Figure 8. Bias of $\hat{a}_i$ for Condition 10

SE. Significant effects on the SE of $\hat{a}_i$ are presented in Table 13. The three-way interaction among model, dual testlet correlation and double-coded items is statistically significant with a small effect size ($\eta_p^2 = 0.015$). This three-way interaction is presented in Figure 9. For the proposed model, the variability among SEs obtained at different levels of dual testlet correlation remains stable across levels of percentage of double-coded items, and the largest SE of $\hat{a}_i$ is always obtained when the dual testlet correlation is small, across levels of the percentage of double-coded items. For models ignoring the double-coded structure (i.e. DTM and 2PL), the variability among SEs of $\hat{a}_i$ obtained at levels of dual testlet correlation changes more when the percentage of double-coded items varies. Models ignoring dual testlets (i.e. T-MIRID, MIRID and 2PL) obtained smallest SE of $\hat{a}_i$ when the dual testlet correlation is 0.5 in a test with 20% of double-coded items. As the percentage of double-coded items increases, the smallest SE of $\hat{a}_i$ yields when the dual testlets are less correlated for T-MIRID, MIRID and 2PL.

Table 13.
ANOVA Results of Significant Effects on the SE of $\hat{a}_i$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	321.232	<0.001	0.381
model * dbcorr	13.231	<0.001	0.048
model * percent_dbcd	6.457	<0.001	0.024
model * dbcorr * percent_dbcd	2.053	0.043	0.015

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

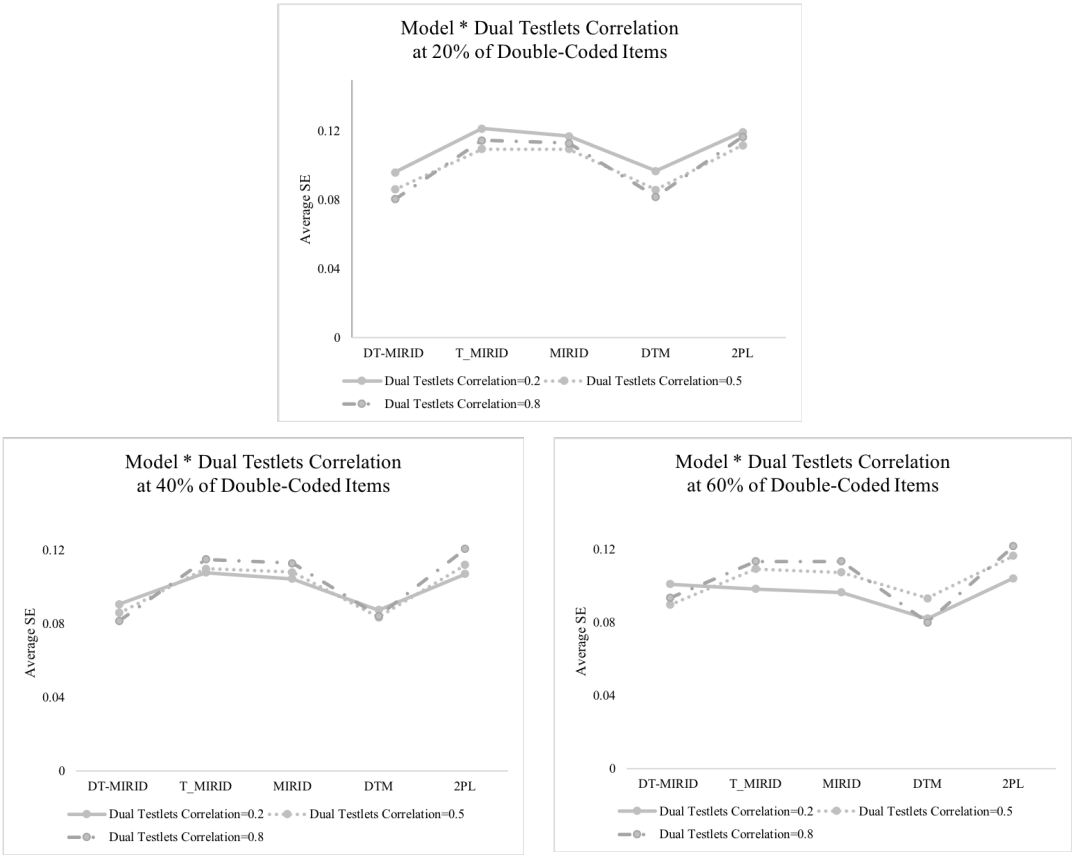


Figure 9. Significant three-way interactions on SE of $\hat{a}_i$— model* dual testlets correlation* percentage of double-coded items

In addition, the results of ANOVA show that the two-way interaction effect between model and dual testlet correlation ($\eta_p^2 = 0.048$) and that between model and percentage of double-coded items ($\eta_p^2 = 0.024$) are statistically significant, each with a small effect size. In addition, the main effect of model is significant on the SE of $\hat{a}_i$

with a large effect size ($\eta_p^2 = 0.381$). An item-level investigation is conducted for the SE of $\hat{a}_i$, but it finds no difference between items nested within a single testlet and items require information from two testlets.

RMSE. Significant effects on the RMSE of $\hat{a}_i$ identified in ANOVA are presented in Table 14. The three-way interaction among model, testlet effect SD and dual testlet correlation is statistically significant with a small effect ($\eta_p^2 = 0.023$). Figure 10 depicts this three-way interaction effect on RMSE of $\hat{a}_i$. For the DT-MIRID, the variability among RMSEs of $\hat{a}_i$ obtained at levels of dual testlet correlation remains stable when the testlet effect SD changes. When the testlet effect SD is small, the DT-MIRID obtains smaller RMSE of $\hat{a}_i$ when the dual testlet correlation is 0.5; when the testlet effect variability is large, the smallest RMSE of $\hat{a}_i$ is obtained with less correlated dual testlet effects (dual testlet correlation=0.2). For models ignoring dual testlet structure (i.e. T-MIRID, MIRID and 2PL), when the testlet effect variability is smaller (i.e. testlet effect SD=0.5), the RMSEs of $\hat{a}_i$ are similar across levels of dual testlets correlation, yet the variability increases when the testlet effect SD becomes larger. This means that the impact of the dual testlets correlation on the RMSEs of $\hat{a}_i$ produced by T-MIRID, MIRID and 2PL is small when the testlet effect SD is small, and the impact is large when the testlet effect is large. For DTM that ignores double-coded items, the variability among RMSEs of $\hat{a}_i$ at levels of the dual testlet correlation is larger when the testlet effect SD is smaller. Moreover, when testlet effect SD is small, the RMSEs produced by different models are more similar than those produced when the testlet effect SD is large, meaning

model mis-specification leads to more error in $\hat{a}_i$ when the testlet effect varability is large.

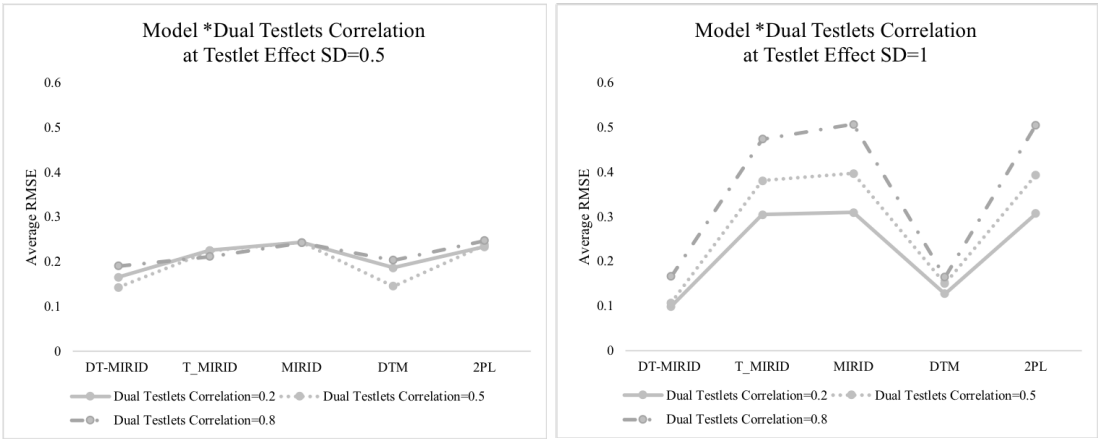


Figure 10. Significant three-way interaction effect on RMSE of $\hat{a}_i$—model* dual testlets correlation* testlet effect SD

Table 14
ANOVA Results of Significant Effects on the RMSE of $\hat{a}_i$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	265.131	<0.001	0.337
model * testlet.sd	100.102	<0.001	0.161
model * dbcorr	3.791	0.015	0.014
model * testlet.sd * dbcorr	6.019	0.001	0.023
Between			
testlet.sd	47.173	<0.001	0.083
dbcorr	12.065	<0.001	0.044
testlet.sd * dbcorr	9.108	<0.001	0.034

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Besides, the interaction effect between model and testlet effect SD ($\eta_p^2 = 0.161$), that between model and dual testlet correlation ($\eta_p^2 = 0.014$), and that between testlet effect SD and the dual testlet correlation ($\eta_p^2 = 0.034$) have significant impact on on RMSE of $\hat{a}_i$, with large, small and small effect sizes, repectively. The results of ANOVA also identify that the main effects of model ($\eta_p^2 =$

0.337), testlet effect SD ($\eta_p^2 = 0.083$) and dual testlets correlation ($\eta_p^2 = 0.044$) are significant effects on the RMSE of $\hat{a}_i$ with large, medium and small effect sizes, respectively.

Similar to the pattern identified in the bias of $\hat{a}_i$, the RMSE of $\hat{a}_i$ is also much larger for an item that requires information from the paired testlets when the dual testlet structure is ignored, compared with items nested within a single testlet. Figure 11 shows the RMSE of $\hat{a}_i$ for each item in condition 10. (Item 1-10 are nested within the first testlet, item 10-21 are nested within the second testlet, and item 21-30 are for paired-testlets.)

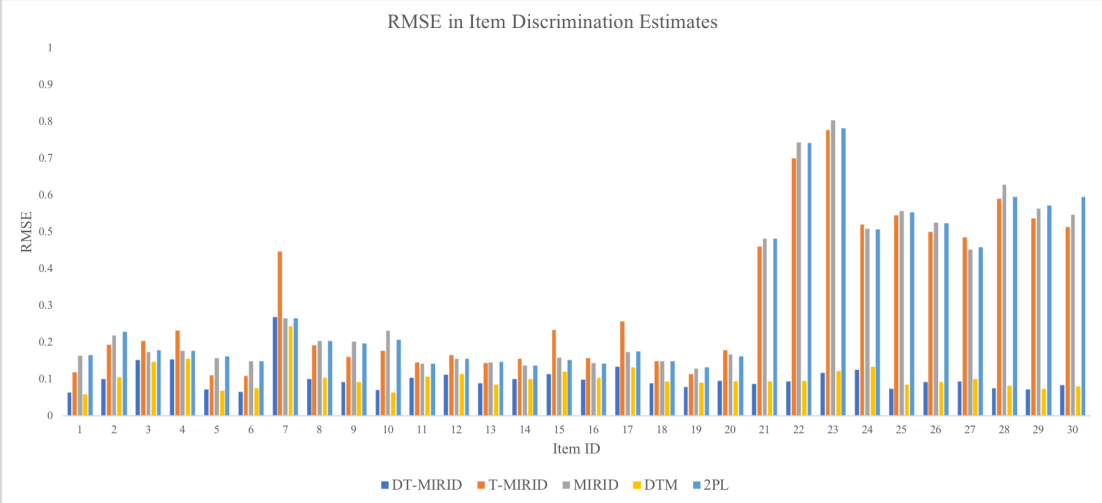


Figure 11. RMSE of $\hat{a}_i$ for Condition 10

Item Difficulty

Due to the issue of unequal cell sample size for item difficulty described in Chapter 3, an evaluation of the homogeneity assumption is used as a screener to decide if ANOVA is to be conducted for error measures of item difficulty parameters. As a result, ANOVA was conducted for the bias and the RMSE of $\hat{b}_i$, but not for SEs.

The results for SEs of $\hat{b}_i$ are summarized by analyzing marginal means of the error measures for each manipulated factor.

Bias. Based on the ANOVA results, the effect of model is statistically significant ($F(1.114, 340.819) = 84.727, p < 0.001$) on the bias of $\hat{b}_i$ with large effect size ($\eta_p^2 = 0.217$). Figure 12 shows the average bias of $\hat{b}_i$ across all simulation conditions for each model. The average bias for $\hat{b}_i$ produced by the proposed model is the smallest among the competing models. The results from the pairwise comparison show significant mean differences of bias for $\hat{b}_i$ among almost all pairs of compared models, except the difference between bias of $\hat{b}_i$ obtained by DTMIRID and those by DTM, and the difference between bias of $\hat{b}_i$ obtained by T-MIRID and those obtained by 2PL. This indicates that ignoring the double-coded item structure does not have a significant impact on the bias of $\hat{b}_i$ for single-coded items.

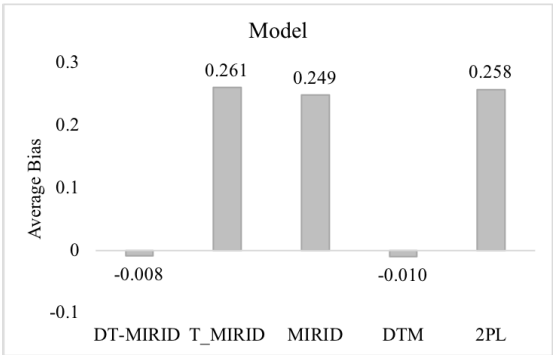


Figure 12. Significant main effect on the bias of $\hat{b}_i$

As observed in bias of $\hat{a}_i$, the $\hat{b}_i$'s for items that require information from dual testlets are also positively biased in a much larger magnitude when the dual testlet structure is ignored, comparing with items nested within a single testlet. Figure 13 presents the bias of $\hat{b}_i$ for all items in condition 10.

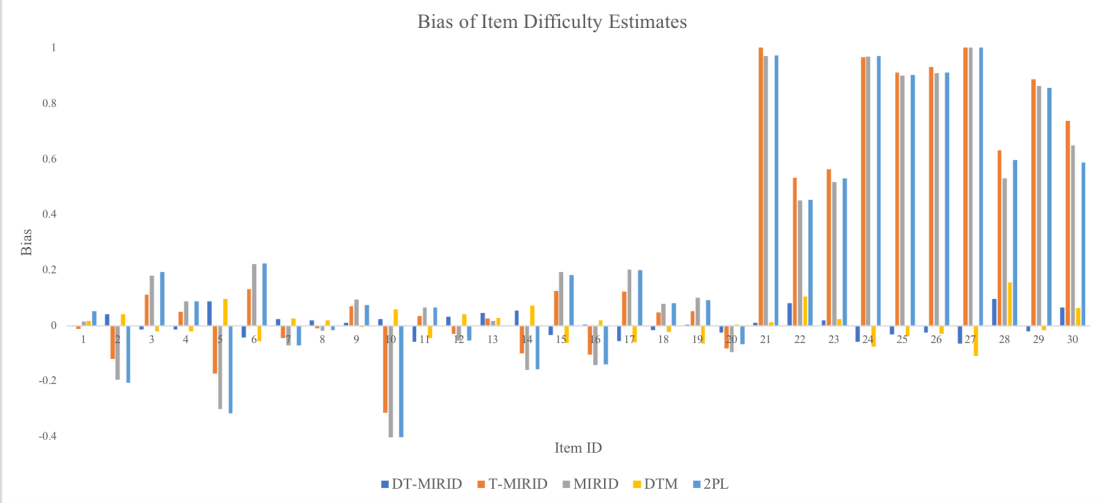


Figure 13. Bias of $\hat{b}_i$ for Condition 10

SE. Judging from the marginal averages, the average of SE for $\hat{b}_i$ decreases when the variability of testlet effects increases, the dual testlets are highly correlated, and there are more double-coded items in the test (See in Figure 14). Smaller SEs of item difficulty are associated with models ignoring dual testlet structure (i.e. T-MIRID, MIRID and 2PL).

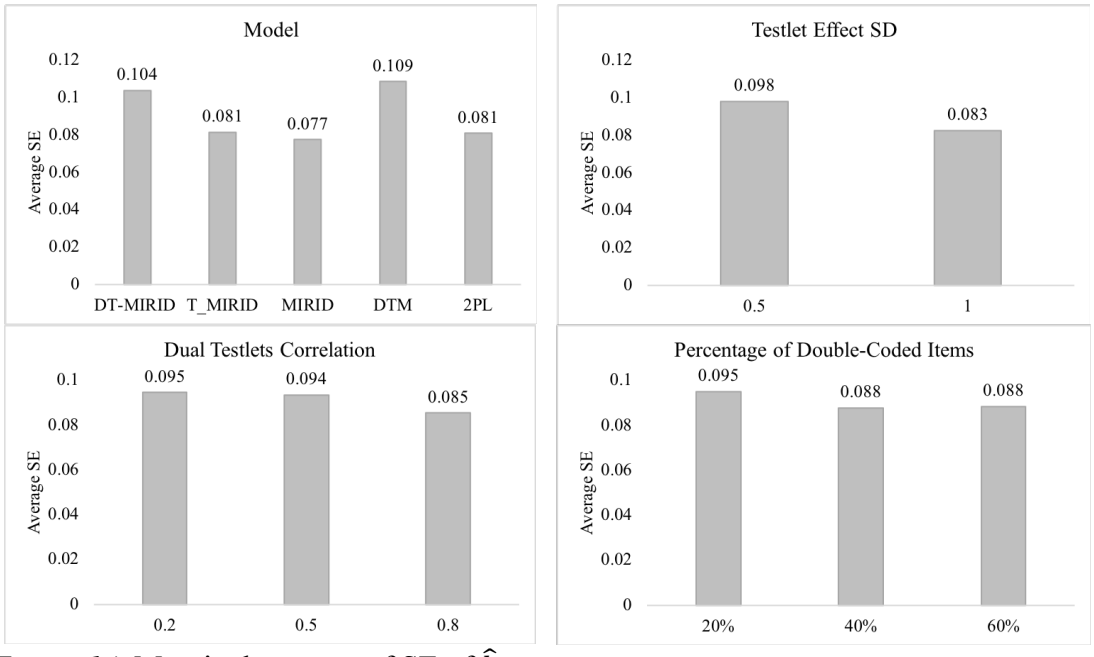


Figure 14. Marginal average of SE of $\hat{b}_i$

The impact of the manipulated factors on the SE of $\hat{b}_i$ for estimates obtained by each model is depicted in Figure 15. The pattern found in the marginal averages of SEs at the levels of testlet effect SD and that found at different levels of dual testlet correlation are consistent across models— the difference between the average of SEs for $\hat{b}_i$ at different levels of testlet effect SD and that at different levels of dual testlet correlation is smaller for DT-MIRID and DTM than those for other models. For models ignoring dual testlet structure (i.e. T-MIRID, MIRID and 2PL), smaller SE of $\hat{b}_i$ is obtained when the percentage of double-coded items is higher. The reversed pattern is observed for estimates yielded by DT-MIRID and DTM— the higher the percentage of double-coded items, the higher the average SE of $\hat{b}_i$.

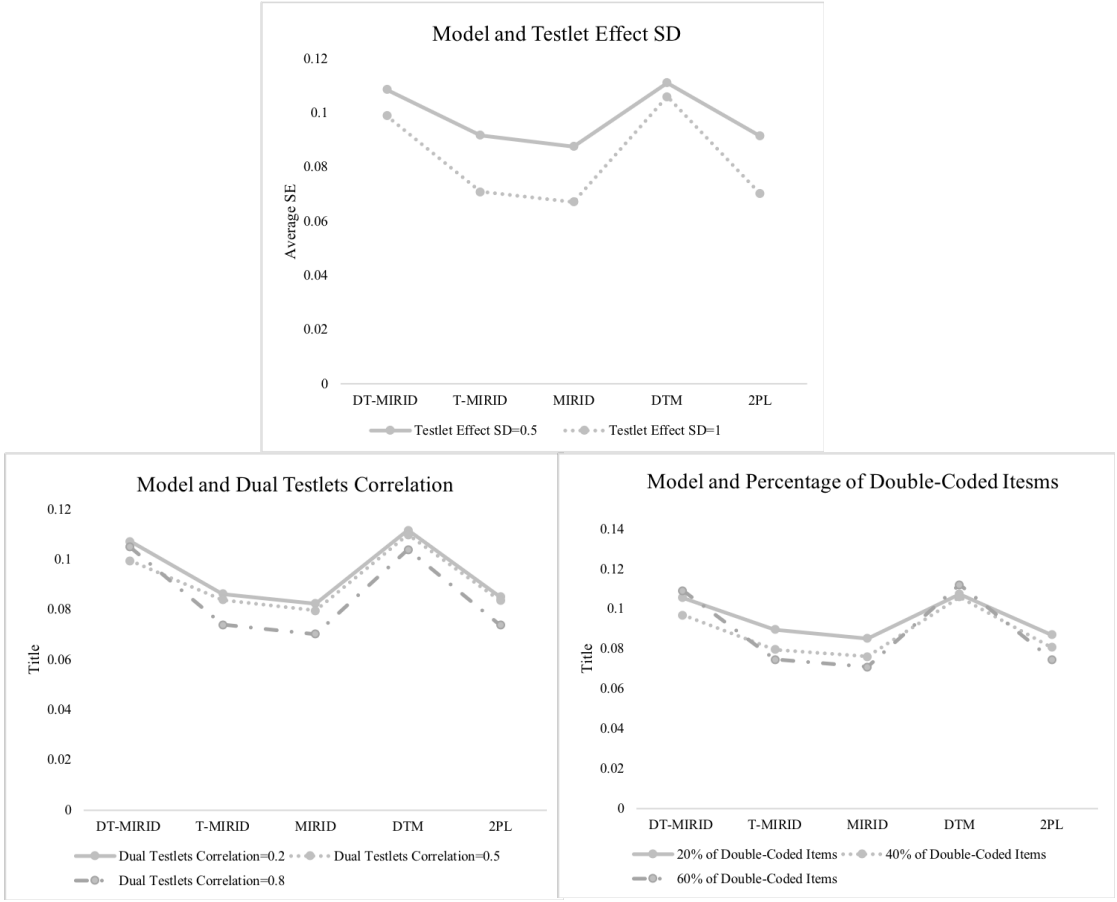


Figure 15. Mean plots of SEs of $\hat{b}_i$

RMSE. The ANOVA results suggest that the main effect of model is the only effect that is statistically significant ($F(1.111, 340.036) = 81.874, p < 0.001$). The effect of model on the RMSE of $\hat{b}_i$ is large ($\eta_p^2 = 0.211$). Pairwise comparisons are conducted to locate the difference between means of RMSEs for $\hat{b}_i$ yielded from pairs of the models. The results of the pairwise comparisons show that the RMSE of $\hat{b}_i$ yielded from the T-MIRID, MIRID and 2PL are very similar. The marginal averages of RMSEs of $\hat{b}_i$ have been presented in Figure 16. The smallest average RMSE of $\hat{b}_i$ is obtained by DT-MIRID. Models accommodating dual testlets (i.e. DT-MIRID and DTM) yield smaller average RMSEs for $\hat{b}_i$, compared to models ignoring dual testlets structure.

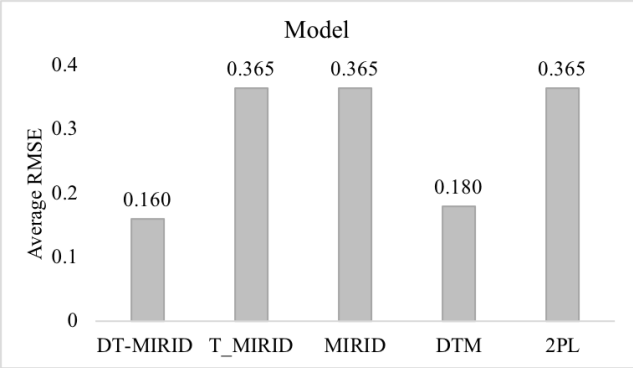


Figure 16 Significant main effect of RMSE of $\hat{b}_i$

The item-level investigation of RMSE of $\hat{b}_i$ has shown that the RMSE for items that require information from the paired testlets are much larger than that for items that are embedded in a single testlet, when the dual testlet structure is ignored. (See pattern in Figure 17)

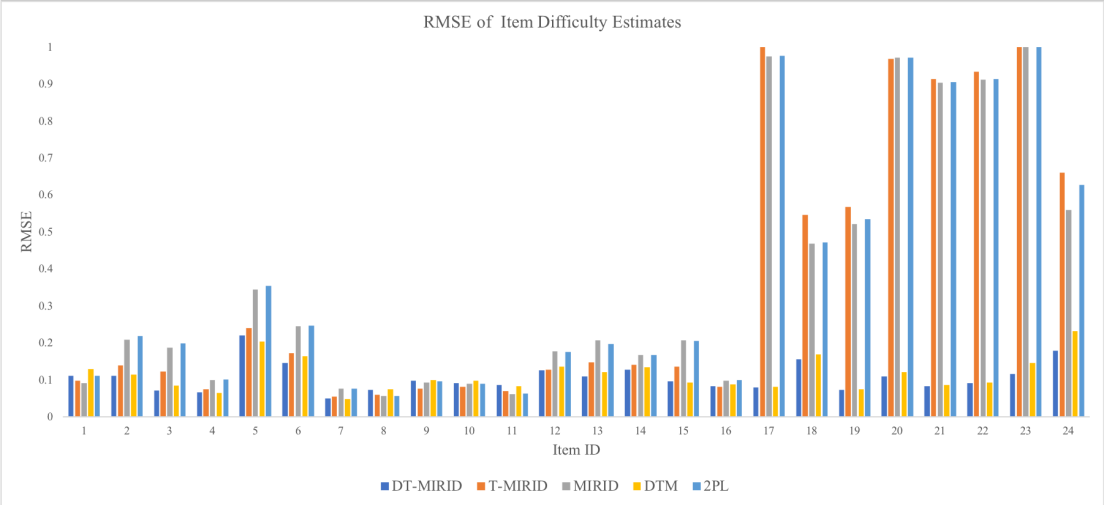


Figure 17. RMSE of $\hat{b}_i$ for Condition 10

Task Weights

As indicated in Table 10, three models accommodate the double-coded item structure—the DT-MIRID, the T-MIRID and the MIRID. Hence, task weight estimates, denoted as $\hat{\sigma}_k$, produced by DT-MIRID, T-MIRID and MIRID are compared.

Bias. In order to understand the influence of the manipulated factors on $\hat{\sigma}_k$ estimated by different models, the mean plots are generated based on the average of bias for $\hat{\sigma}_k$ at all levels of each manipulated factor for each model (See Figure 18). Small average bias of $\hat{\sigma}_k$ is obtained when the variability of testlet effect is large and the dual testlets are highly correlated. Yet, the $\hat{\sigma}_k$ estimated by DT-MIRID is less impacted by the change of testlet effect variability as the difference between the averages of bias of $\hat{\sigma}_k$ at different levels of testlet effect SD is much smaller than that produced by models ignoring dual testlet structure (i.e. T-MIRID and MIRID). Besides, the bias of $\hat{\sigma}_k$ produced by the DT-MIRID is negative and more stable than those produced by T-MIRID and those produced by MIRID when the dual testlets are

weakly or moderately correlated. When DT-MIRID is used to obtain the $\hat{\sigma}_k$, the smallest bias is obtained when there are 60% of double-coded items. For T-MIRID and MIRID, the lowest bias of $\hat{\sigma}_k$ is obtained when only 20% of double-coded items are in the test. Whereas, the lowest average bias of $\hat{\sigma}_k$ is obtained when there are 60% double-coded items in the test for DT-MIRID.

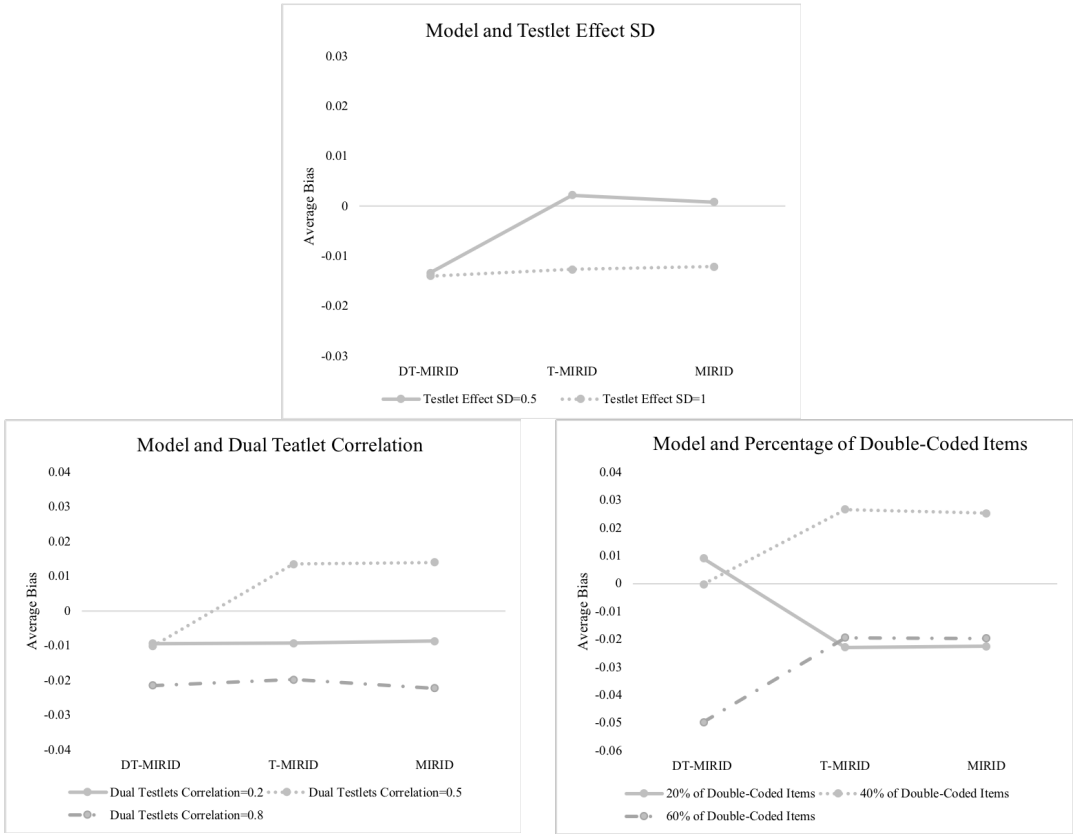


Figure 18. Mean plots of bias for $\hat{\sigma}_k$

SE. Generally speaking, the DT-MIRID, T-MIRID and MIRID yield very similar SEs of $\hat{\sigma}_k$ across study conditions. The smallest average SE for $\hat{\sigma}_k$ across all study conditions is produced by the MIRID. In addition, when the variability of testlet effects is larger, the study obtains smaller SEs of $\hat{\sigma}_k$. Moreover, small SEs of $\hat{\sigma}_k$ are obtained when the correlation between testlet effects from the dual testlets is 0.5 for all three models. Small average SE of $\hat{\sigma}_k$ also tends to be associated with a higher

percentage of double-coded items. Figure 19 presents the mean plots of SE of $\hat{\sigma}_k$ for manipulated factors.

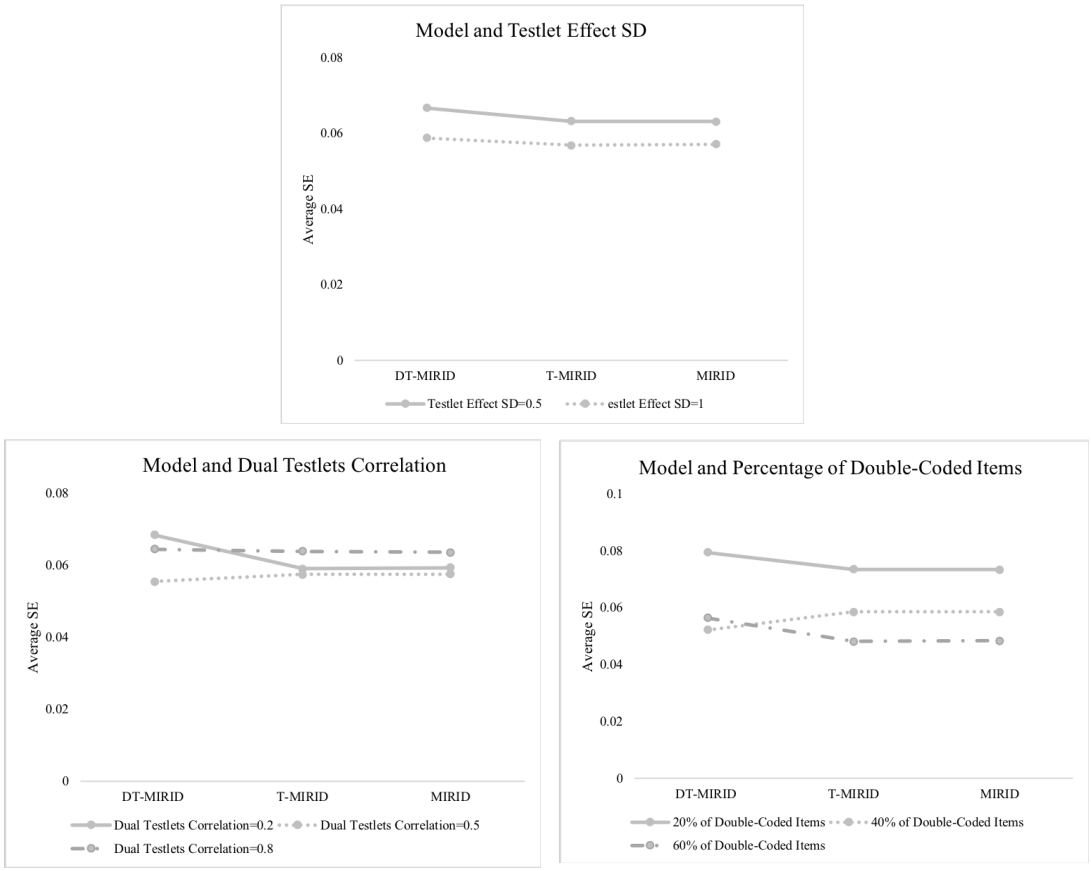


Figure 19. Mean plots of SE for $\hat{\sigma}_k$

RMSE. For the DT-MIRID, a smaller average of RMSEs for $\hat{\sigma}_k$ is obtained when the testlet effect SD is larger. Whereas in TMIRID and MIRID, the pattern is reversed—smaller means of RMSEs for $\hat{\sigma}_k$ is associated with smaller testlet effect SD. This indicates that there is more error in the task weight estimates for models ignoring dual testlets than that for DT-MIRID when the testlet effect SD is large. The smallest average RMSEs of $\hat{\sigma}_k$ are produced when the dual testlets correlation is 0.2 for all models. The average of RMSEs of $\hat{\sigma}_k$ yielded from DT-MIRID is less sensitive to the change of dual testlet correlation. In terms of the percentage of double-coded items, the largest mean of RMSEs of $\hat{\sigma}_k$ is always obtained when there are 60% of

double-coded items for the three models, indicating weight estimates tend to contain more error with a test that contains a large proportion of double-coded items. Figure 20 presents the mean plots of RMSE for $\hat{\sigma}_k$ at levels of the manipulated factors for each model.

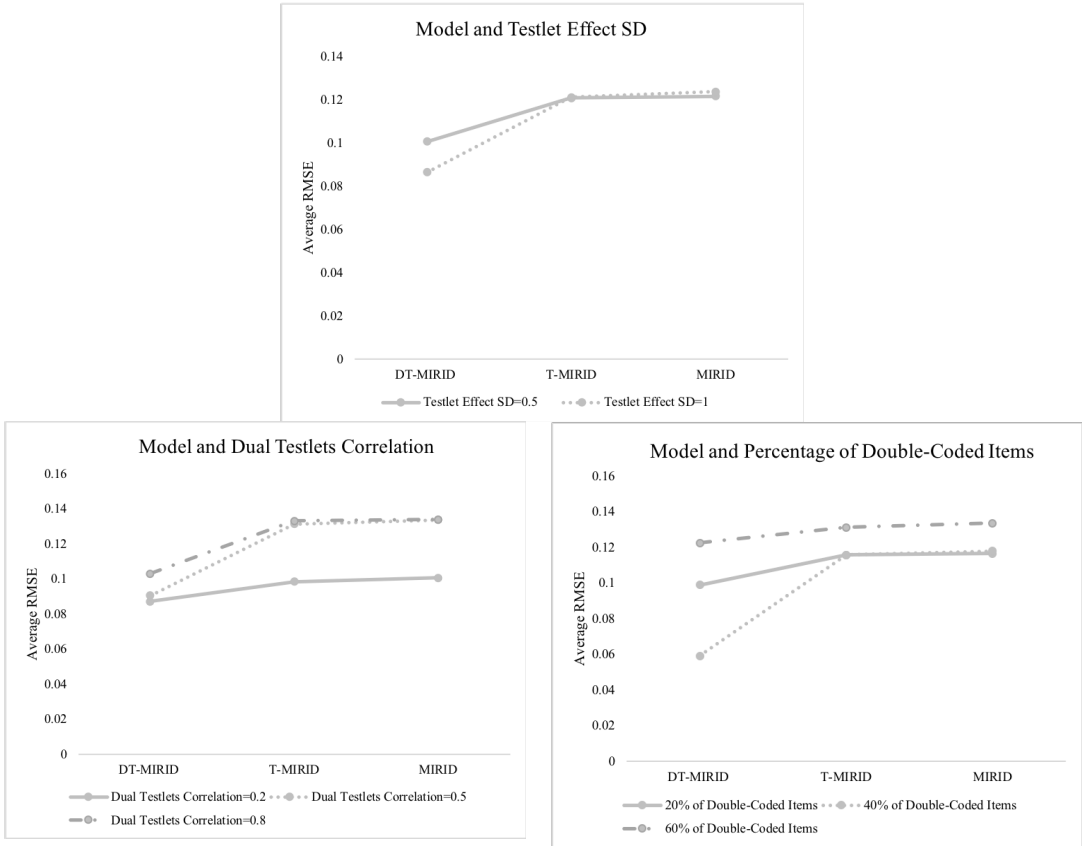


Figure 20. Mean plots of RMSE for $\hat{\sigma}_k$

Intercept

The same as that for task weight, the intercept (denoted as τ) is only included in a model where the double-coded item structure is correctly specified. Therefore, $\hat{\tau}$ estimated using DT-MIRID, T-MIRID and MIRID are compared in this study.

Bias. The mean plots of the bias for $\hat{\tau}$ is generated for all manipulated factors with each model (See in Figure 21). When the variability of the testlet SD is 0.5, the intercept estimate is positively biased; and when the testlet SD is 1, the intercept

estimate is negatively biased. Comparing with T-MIRID and MIRID, the DT-MIRID yields the smallest difference between the bias of $\hat{\tau}$ obtained at the lower level of testlet effect SD and that obtained at the higher level of testlet effect SD. Similarly, bias of $\hat{\tau}$ produced by DT-MIRID are less impacted by the magnitude of dual testlet correlation. The smallest absolute average bias of $\hat{\tau}$ is obtained for all the three models when the dual testlet correlation is 0.2. It also shows that $\hat{\tau}$ is less biased when there are more double-coded items.

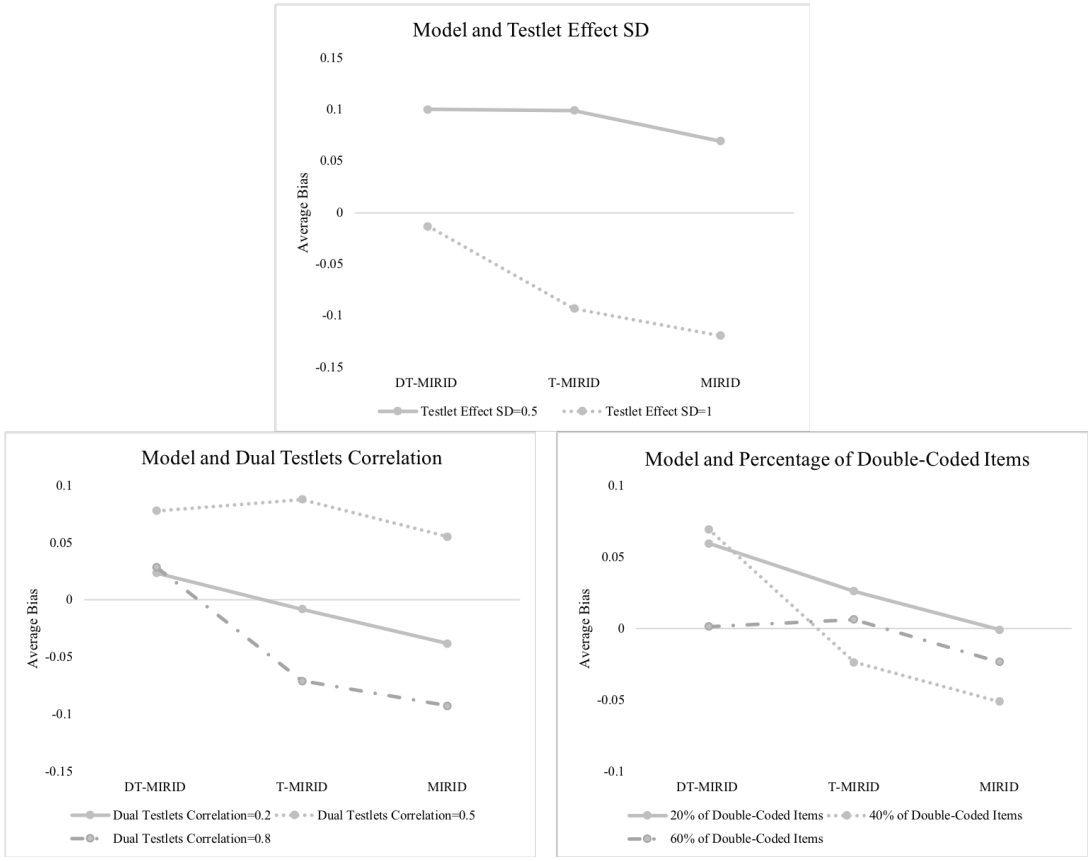


Figure 21. Mean plots of bias for $\hat{\tau}$

SE. The SEs of $\hat{\tau}$ does not differ much across study conditions and across competing models. They range from 0.022 to 0.010. In general, smaller SEs of the intercept estimate are identified when the testlet effect SD is larger, the dual testlets are less correlated, and the test contains more double-coded items. These patterns are

consistent across competing models. The SE of $\hat{\tau}$ produced by the DT-MIRID is the least sensitive among the three models toward the change of testlet effect SD but is more impacted by the change of dual testlet correlation.

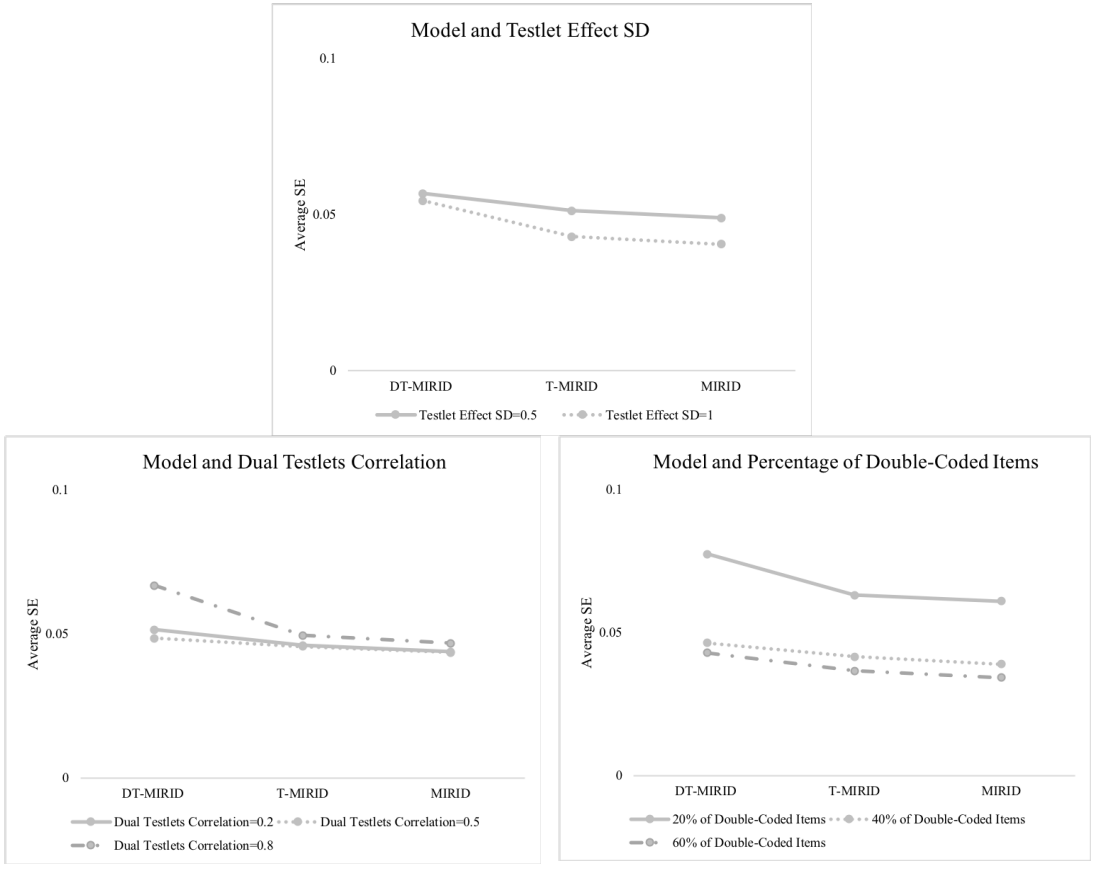


Figure 22. Mean plots of SE for $\hat{\tau}$

RMSE. Small averages of RMSEs of $\hat{\tau}$ is obtained when the testlet effect SD is 1 for DT-MIRID and T-MIRID. Whereas MIRID, which ignores all testlet structure, produces the smallest average of RMSEs of $\hat{\tau}$ when the variability of testlet effect is small. In terms of RMSEs of $\hat{\tau}$ at different levels of percentage of double-coded items, the DT-MIRID produces the most stable estimates when the percentage of double-coded item changes. In addition, the RMSEs of $\hat{\tau}$ yielded from the three models are very similar when the dual testlets correlation is 0.2. In other words, the

impact of ignoring dual testlet structure is minimal on the estimated intercept when the dual testlets are less correlated.

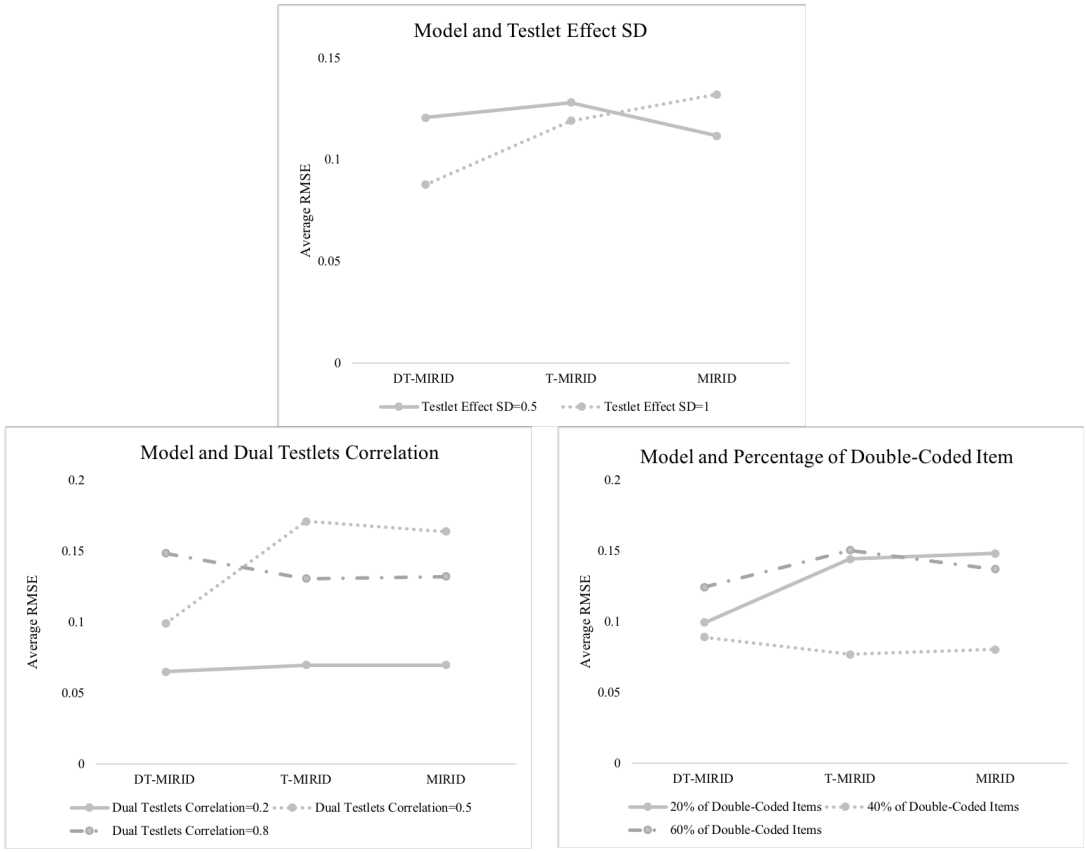


Figure 23. Mean plots of RMSE for $\hat{\tau}$

Testlet Effect Variance

The testlet effect variances of the first and the second testlets in the test structure is compared among the DT-MIRID, the T-MIRID and the DTM. As described in Chapter 3, the testlet effect variances (denoted as VAR_γ) are generated to be equal for the two testlets.

Bias. At all study conditions, models accommodating the dual testlet structure (i.e. DT-MIRID and DTM) overestimate testlet variance, whereas the model ignoring the dual testlet structure underestimates testlet variance. The bias of $\widehat{VAR_\gamma}$ obtained by DT-MIRID and the DTM are more stable than those obtained by T-MIRID when

the true testlet effect variance changes. Moreover, the $\widehat{VAR_\gamma}$ deviates more from the true value when the dual testlets are highly correlated, for all three models. The bias of $\widehat{VAR_\gamma}$ yielded from the T-MIRID is more stable than that from the DT-MIRID and the DTM to the change in the percentage of double-coded items.

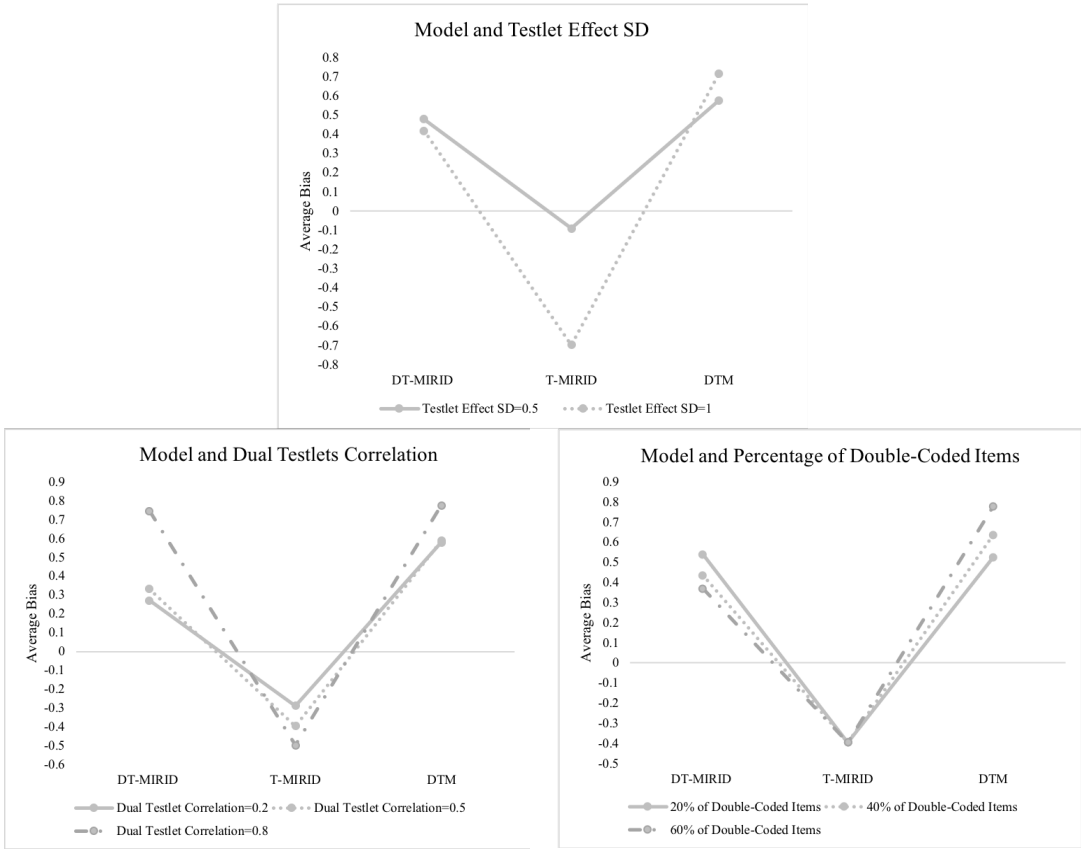


Figure 24. Mean plots of bias for $\widehat{VAR_\gamma}$

SE. The T-MIRID produces the smallest SEs of $\widehat{VAR_\gamma}$ for all study conditions. Besides, the average of SEs for $\widehat{VAR_\gamma}$ yielded from the T-MIRID is less impacted by the change of the manipulated factors than those from the DT-MIRID and DTM. For all models, large average of SEs for $\widehat{VAR_\gamma}$ is always produced when the testlet effect variability is large and when there are 60% of double-coded items.

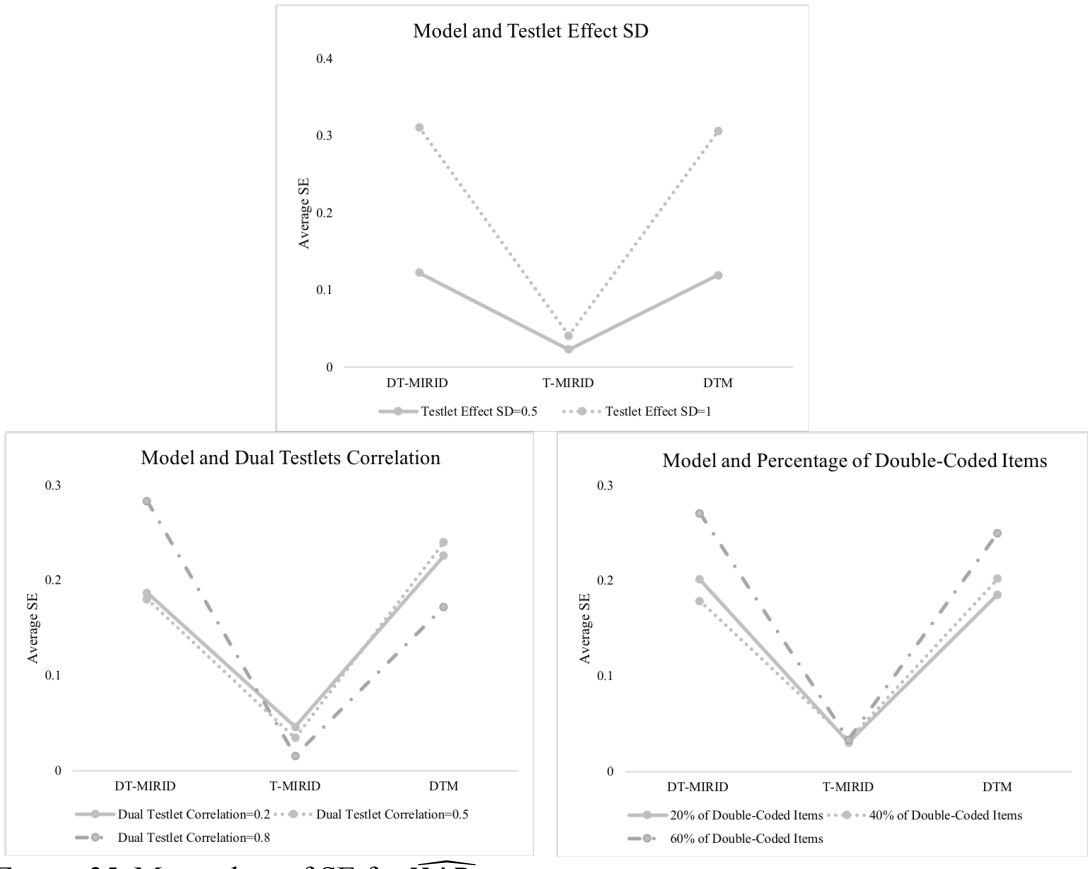


Figure 25. Mean plots of SE for $\widehat{VAR_\gamma}$

RMSE. In spite of the model used in estimating the testlet effect variances, the average RMSEs of $\widehat{VAR_\gamma}$ is always the largest when the dual testlets correlation is 0.8 and the true testlet effect variance is 1. The average of RMSEs of $\widehat{VAR_\gamma}$ is more stable for estimates produced by DT-MIRID and DTM when the true testlet effect variability changes. Whereas the average of RMSEs for $\widehat{VAR_\gamma}$ produced by T-MIRID increases dramatically when the true testlet effect variability increases. The average of RMSEs for $\widehat{VAR_\gamma}$ estimated by DTM, which ignores the double-coded item structure, is more impacted by the change in the percentage of double-coded items. The DTM produces larger average of RMSEs for $\widehat{VAR_\gamma}$, when the test contains more double-coded items.

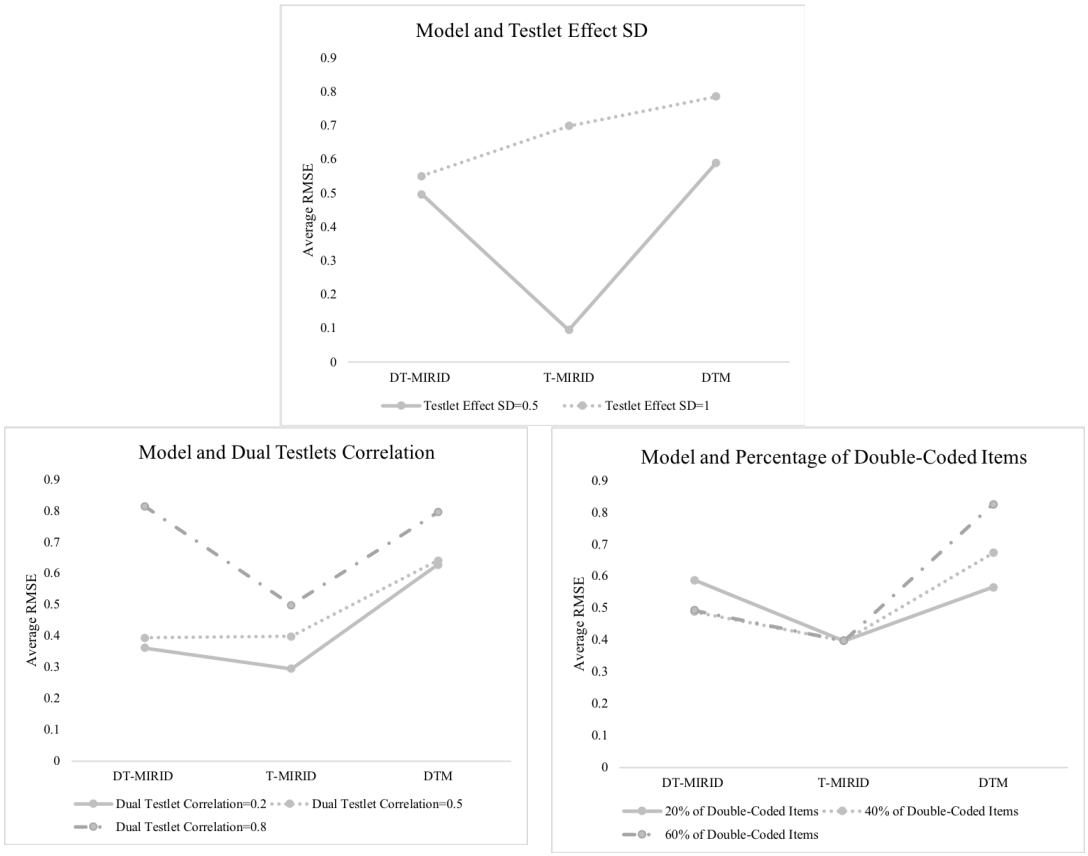


Figure 26. Mean plots of RMSE for $\widehat{VAR_\gamma}$

Correlation between Testlet Effects of Dual Testlets

The correlation between the dual testlets (denoted as $\rho_{\gamma_1\gamma_2}$) is only estimated in DT-MIRID and DTM. Therefore, the error measures of $\hat{\rho}_{\gamma_1\gamma_2}$ are compared between DT-MIRID and DTM at levels of the manipulated factors.

Bias. For the estimates produced by both DT-MIRID and DTM, smaller values of bias for $\hat{\rho}_{\gamma_1\gamma_2}$ are obtained when the true testlet effect SD is larger and the true correlation between testlet effects for the dual testlets is larger. That is, when the true testlet effect variability is larger and the dual testlets are highly correlated, the $\hat{\rho}_{\gamma_1\gamma_2}$ is estimated with less bias. In addition, the fact that DTM does not model the

double-coded items leads to more biased $\hat{\rho}_{\gamma_1\gamma_2}$, especially in a test that contains larger proportion of double-coded items.

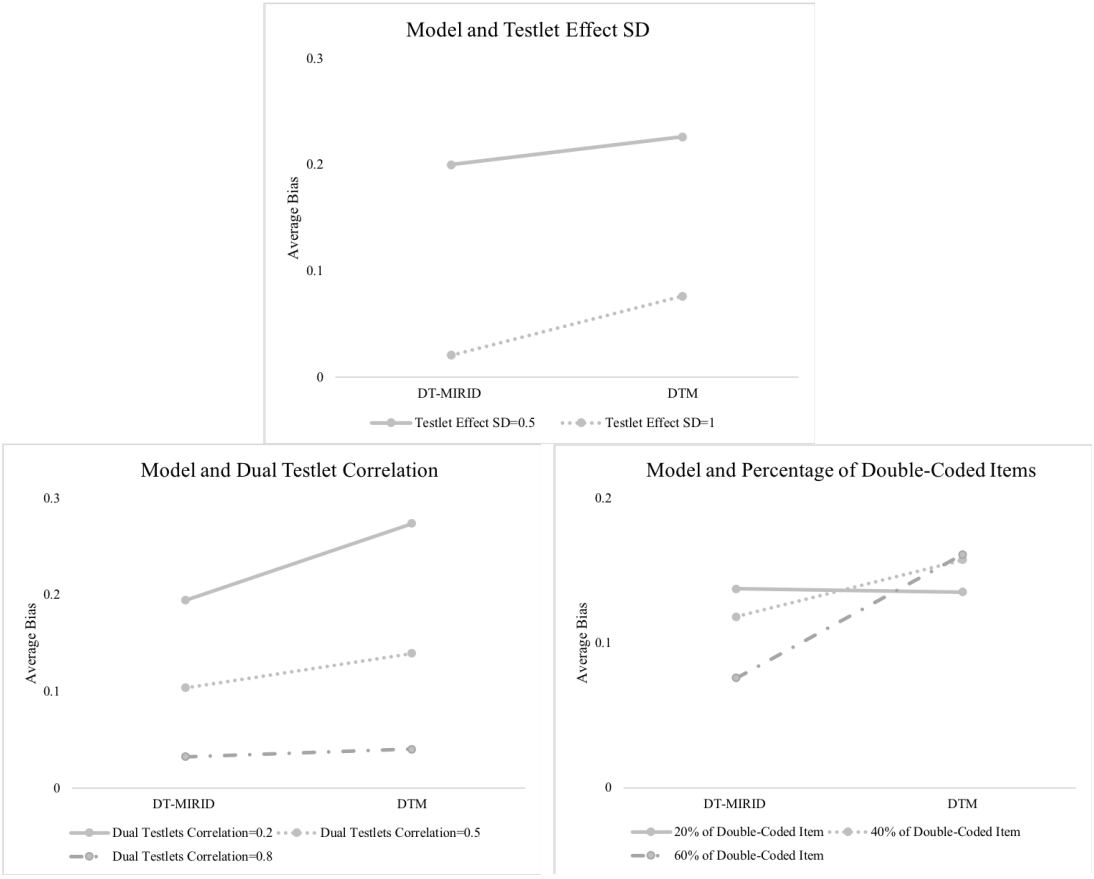


Figure 27. Mean plots of bias for $\hat{\rho}_{\gamma_1\gamma_2}$

SE. In most of the study conditions, the SE of $\hat{\rho}_{\gamma_1\gamma_2}$ is larger for $\hat{\rho}_{\gamma_1\gamma_2}$ estimated by DT-MIRID. Similar to what have been found in the bias of $\hat{\rho}_{\gamma_1\gamma_2}$, the SEs of $\hat{\rho}_{\gamma_1\gamma_2}$ tend to be smaller when the true testlet effect SD is larger and the testlet effects from the dual testlets are more correlated. For DTM, the SE of $\hat{\rho}_{\gamma_1\gamma_2}$ is not heavily impacted by the percentage of double-coded items in the test. Whereas the SEs of $\hat{\rho}_{\gamma_1\gamma_2}$ from the DT-MIRID is less stable across replications when the percentage of double-coded items changes—more double-coded in the test leads to more stable estimation of $\hat{\rho}_{\gamma_1\gamma_2}$.

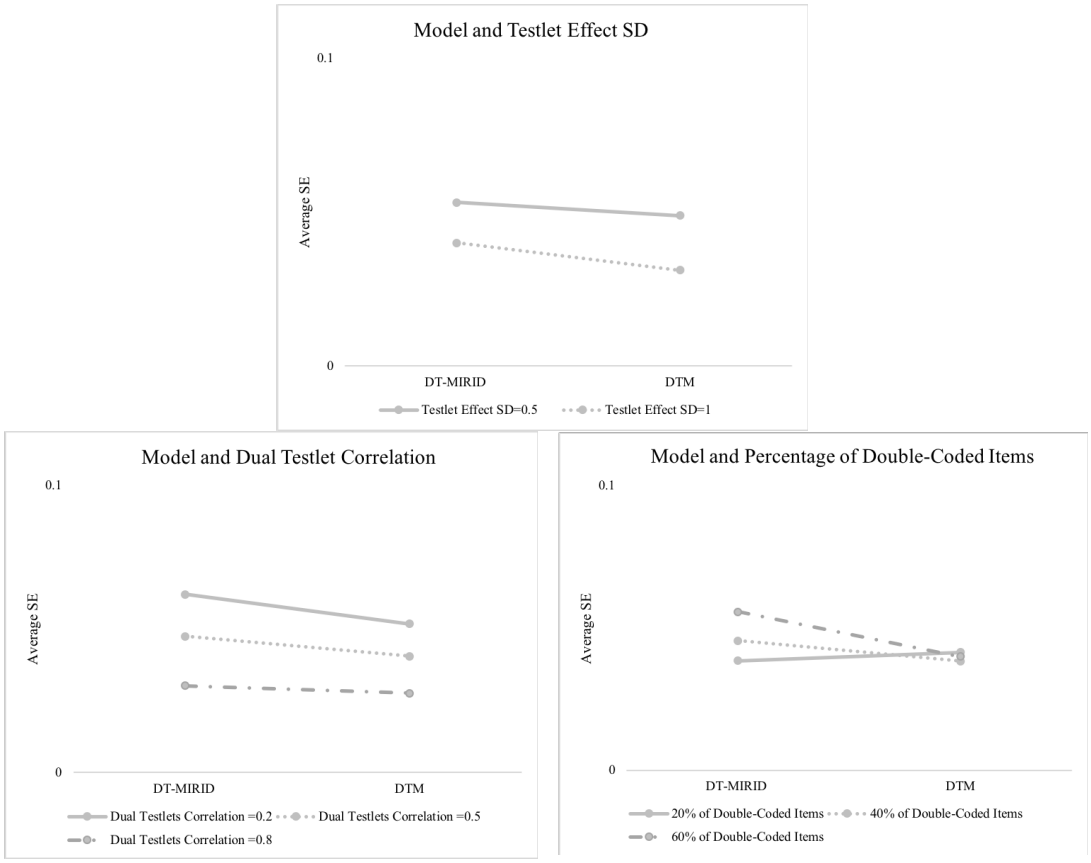


Figure 28. Mean plots of SE for $\hat{\rho}_{\gamma_1\gamma_2}$ for manipulated factors

RMSE. In terms of the total error in $\hat{\rho}_{\gamma_1\gamma_2}$, the RMSE showed consistent patterns with what have been found in bias and SEs— (a) the DT-MIRID produces less total error than the DTM in estimating $\hat{\rho}_{\gamma_1\gamma_2}$; (b) the larger the testlet effect variability is, the smaller the RMSEs for $\hat{\rho}_{\gamma_1\gamma_2}$ are; (c) the larger the true dual testlets correlation is, the more accurate the estimates of $\rho_{\gamma_1\gamma_2}$ are; and (d) the impact of ignoring the double-coded item structure on the RMSE of $\hat{\rho}_{\gamma_1\gamma_2}$ is more severe in conditions where there are more double-coded items.

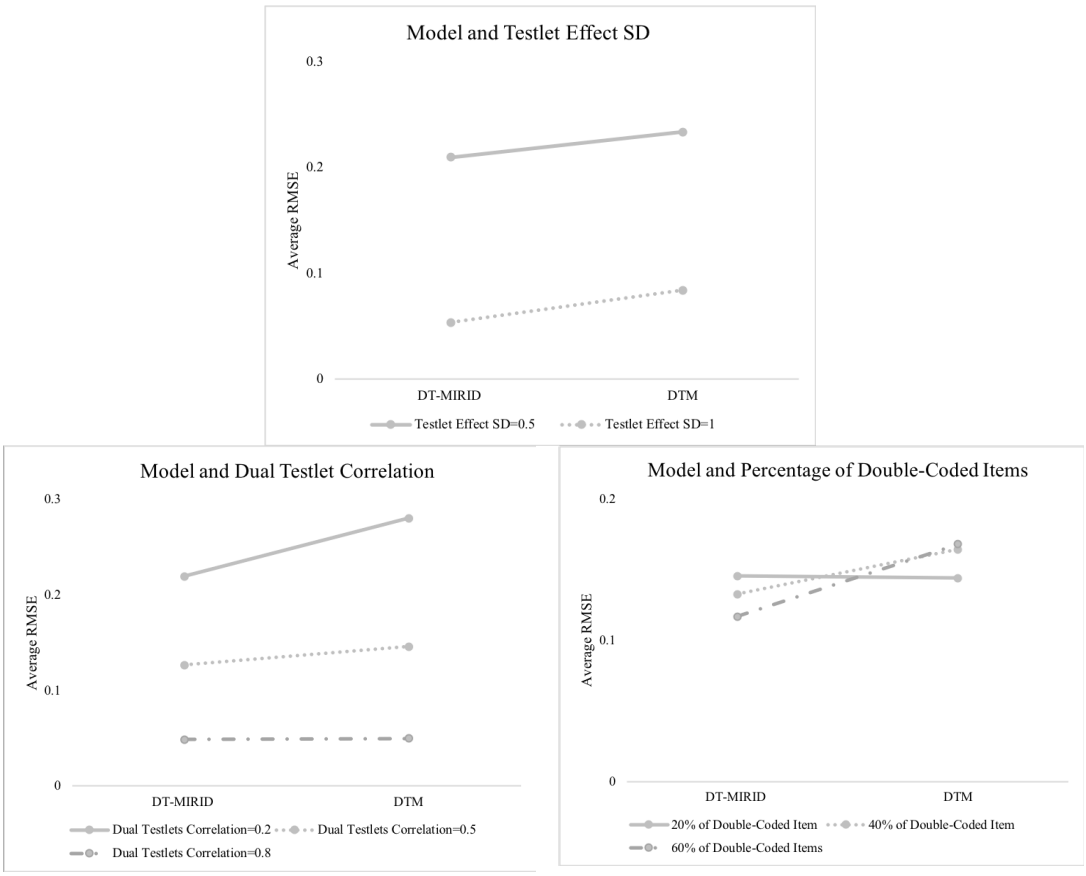


Figure 29. Mean plots of RMSE for $\hat{\rho}_{\gamma_1\gamma_2}$

Overall Ability

Bias. The ANOVA results indicate that no effect on the bias of $\hat{\theta}_j$ is statistically significant with at least a small effect size. This is due to the fact that the mean of the overall ability estimates is constrained to be 0 in estimation to set the scale for the relative location between the ability and the item difficult. The bias of $\hat{\theta}_j$ is, therefore, centered around 0.

The SD of the bias for $\hat{\theta}_j$ is calculated for each study condition (See in Appendix D). Across all study conditions, the NCS yields the largest SD of bias for $\hat{\theta}_j$, whereas the SDs of bias for $\hat{\theta}_j$ from pattern scoring models are very similar. The SD of bias for $\hat{\theta}_j$ increases as the testlet effect SD increases and as the dual testlet

correlation decreases. The SD of bias for $\hat{\theta}_j$ estimated using NCS is also less stable when the testlet effect SD or the dual testlets correlation changes, comparing with patterns scoring models.

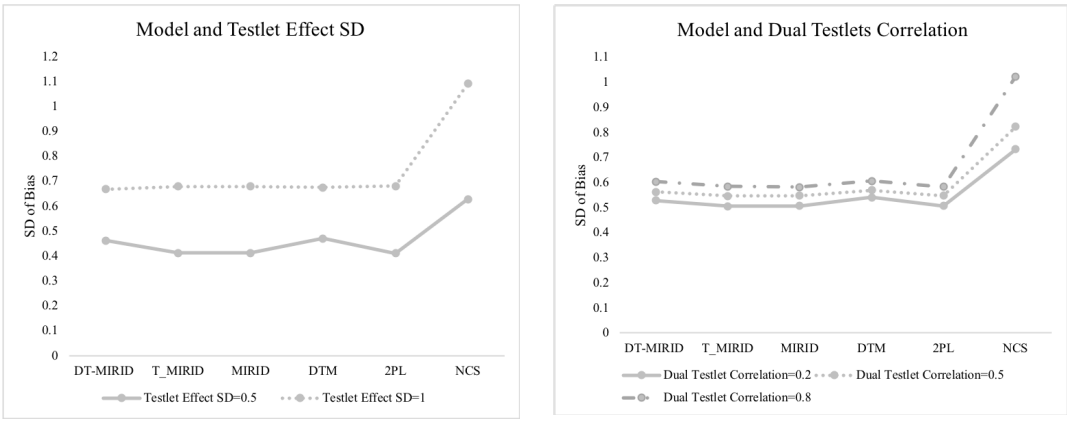


Figure 30. Mean plot of the SD of bias for $\hat{\theta}_j$

SE. The significant effects identified by ANOVA are tabulated in Table 15. Based on the results, the significant three-way interaction effect among model, testlet effect SD and dual testlets correlation has a small effect size ($\eta_p^2 = 0.011$). Figure 31 presents this three-way interaction. Two observations are made— (a) the variability among the averages of SEs for $\hat{\theta}_j$ at different levels of dual testlets correlation is generally smaller when the testlet effect SD is small for all models, meaning that the SE of $\hat{\theta}_j$ tends to be more stable towards the change of dual testlets correlation when the testlet effect SD is small, and (b) the NCS produces the largest SE when the dual testlets correlation is large in spite of the testlet effect SD, whereas the pattern scoring models yield smaller SE of $\hat{\theta}_j$ when the dual testlets correlation is at 0.8.

Table 15

ANOVA Results of Significant Effects on the SE of $\hat{\theta}_j$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	152534.616	<0.001	0.895
model * testlet.sd	636.709	<0.001	0.034
model * dbcorr	1060.767	<0.001	0.106
model * testlet.sd * dbcorr	100.136	<0.001	0.011
Between			
testlet.sd	2337.776	<0.001	0.115
dbcorr	120.696	<0.001	0.013

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

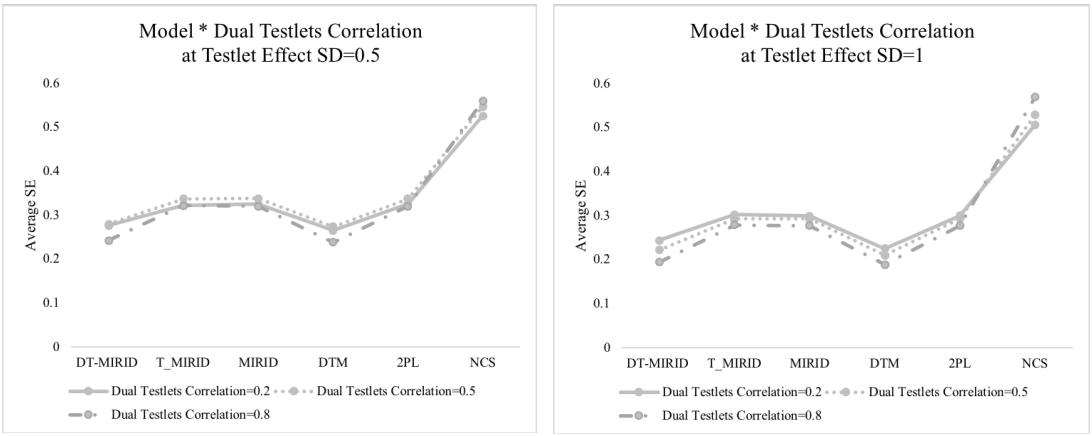


Figure 31. Significant three-way interaction effects on the SE of $\hat{\theta}_j$— model*dual testlets correlation*testlet effect SD

Moreover, the interaction effect between model and testlet effect SD has a small effect on the SEs of $\hat{\theta}_j$ ($\eta_p^2 = 0.034$). The two-way interaction between model and dual testlet correlation has a large effect on the SEs of $\hat{\theta}_j$ ($\eta_p^2 = 0.106$). In addition, the ANOVA also indicates that the main effect of model ($\eta_p^2 = 0.895$), testlet effect SD ($\eta_p^2 = 0.115$) and the dual testlets correlation ($\eta_p^2 = 0.013$) have large, large and small effects on the SEs of $\hat{\theta}_j$, respectively.

RMSE. According to the ANOVA results in Table 16, the interaction effect between model and testlet effect SD ($\eta_p^2 = 0.028$), and that between model and dual

testlet correlation ($\eta_p^2 = 0.03$) both have small effects on the RMSE of $\hat{\theta}_j$. Trends found in these two interactions are similar (see in Figure 32)— (a) both interactions are ordinal; (b) the RMSE of $\hat{\theta}_j$ estimated by NCS are more inflated by the increase of testlet effect SD and by the increase of dual testlets correlation; and (c) among pattern scoring models, models accommodating dual testlet structure (i.e. DT-MIRID and DTM) are less impacted by the change of testlet effect SD and the dual testlets correlation.

Table 16
ANOVA Results of Significant Effects on the RMSE of $\hat{\theta}_j$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	11010.803	<0.001	0.380
model * testlet.sd	518.667	<0.001	0.028
model * dbcorr	274.686	<0.001	0.03
Between			
testlet.sd	1698.735	<0.001	0.086
dbcorr	101.548	<0.001	0.011

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

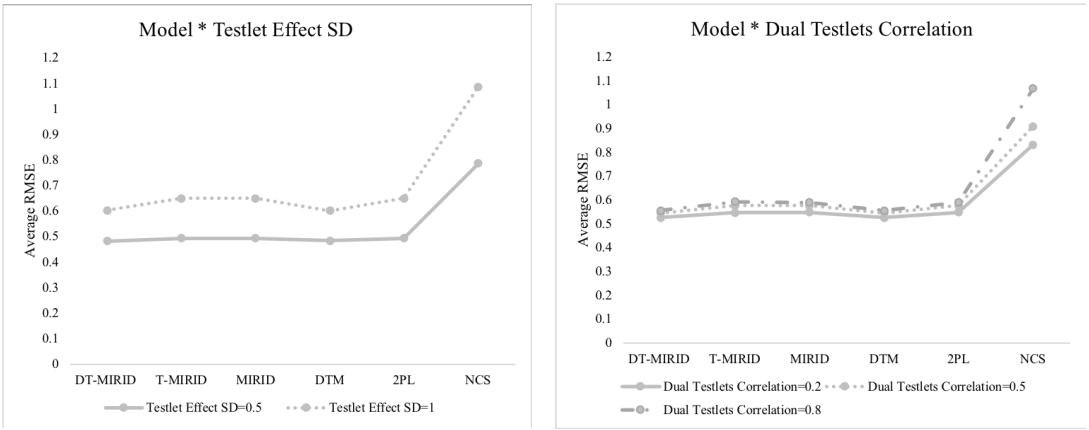


Figure 32. Significant two-way interaction effects on RMSE of $\hat{\theta}_j$

In addition, the main effect of model ($\eta_p^2 = 0.380$), testlet effect SD ($\eta_p^2 = 0.086$) and dual testlet correlation ($\eta_p^2 = 0.011$) are significant on the RMSEs of $\hat{\theta}_j$ with large, medium and small effect, respectively.

Subscore of Multiplication (as an Example of Subscores)

The test that this simulation study is based on contains 4 subscores – subscore of addition, subscore of subtraction, subscore of multiplication and subscore of division. Although these subscores represent students' latent ability in different content domains, the actual content and the difference among content domains are not the focus of this study. As described in Chapter 3, the test blueprint is designed to ensure balanced assessment of the 4 subdomains, the subscore structures are designed to be as similar as possible. Hence, the ANOVA results of the 4 subscores have little difference. That is, the significant effects with at least a small effect size identified in each of the subscores are largely consistent across the 4 subscores. The pattern in marginal averages and interactions are also very similar. Therefore, this section presents only the subscore estimation accuracy for the subscore of multiplication as an example. The subscore of multiplication is selected as an example because its significant effects are also common in other subscores. There are a few effects that are not significant for error measures of subscore for multiplication but significant for error measures of other subscores. Since these effects are not commonly observed significant effects across subscores, they will be briefly summarized after the ANOVA results of subscore for multiplication. The ANOVA results for other subscores are in Appendix E.

Bias. According to the results of ANOVA, the model has a significant effect on the bias of $\hat{\theta}_{jM}$, with a small effect size ($F(1.039,18689.001) = 627.4, p < 0.001, \eta_p^2 = 0.034$). Figure 33 shows small marginal means of bias for $\hat{\theta}_{jM}$ are obtained when 2PL or the pattern scoring models that accommodate the dual testlet structure are used in estimating the subscore of multiplication. All Pairwise comparison demonstrated significant mean bias difference for $\hat{\theta}_{jM}$, except those between T-MIRID and NCS, and those between MIRID and NCS.

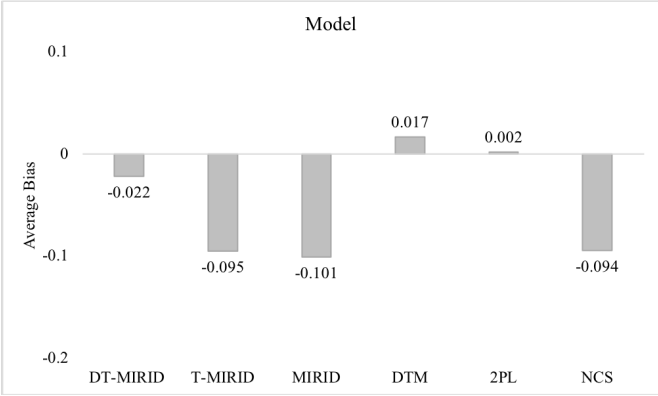


Figure 33 Significant main effect on the bias of $\hat{\theta}_{jM}$

The less informative prior was assumed for $\hat{\theta}_{jM}$ with a normal distribution with a mean of 0 a variance of 4. The mean bias of $\hat{\theta}_{jM}$ is close to 0 for all conditions. The SDs of bias for $\hat{\theta}_{jM}$ is calculated for each study condition and for each model (See in Appendix D). The examination of SDs of bias for $\hat{\theta}_{jM}$ has shown that when the variability of the true testlet effects is large, the SD of bias for $\hat{\theta}_{jM}$ tends to be larger. In addition, the SD of the bias for $\hat{\theta}_{jM}$ is much larger for estimates by NCS than those estimated by pattern scoring models. These patterns on the SDs of bias for subscore estimates are identical across subscores.

SE. The ANOVA results (see in Table 17) show that the three-way interaction effect among model, testlet effect SD and dual testlets correlation is significant with a small effect ($\eta_p^2 = 0.016$). The models ignoring dual testlet structure (i.e. T-MIRID, MIRID and 2PL) are impacted more heavily by the change in dual testlet correlation when testlet effect increases, whereas the SE of $\hat{\theta}_{jM}$ remains stable for DT-MIRID, DTM and NCS at different levels of the dual testlet correlation and when the testlet effect SD changes (See in Figure 34). The three-way interaction among model, testlet effect SD and percentage of double-coded items has a significant impact on the SE of $\hat{\theta}_{jM}$ with small effect size ($\eta_p^2 = 0.023$). Figure 35 demonstrates that the means of SEs for $\hat{\theta}_{jM}$ are less variable at different levels of dual testlet correlations when testlet effect is low, whereas much lower SEs of $\hat{\theta}_{jM}$ are produced for a test with 60% of double-coded items at high level of testlet effect SD for all models, and especially for the NCS estimates. The interaction among model, dual testlets correlation and the percentage of double-coded items are statistically significant with a medium effect size ($\eta_p^2 = 0.094$). Such interaction is depicted in Figure 36, where the impact of dual testlet correlation on the SE of $\hat{\theta}_{jM}$ increases as the percentage of double-coded items increases. The SEs of $\hat{\theta}_{jM}$ at different levels of dual testlet correlation yielded from NCS are most sensitive to the change of dual testlet correlation when there are 40% of double-coded items in the test. The three-way interaction among testlet effect SD, dual testlets correlation and the percentage of double-coded items is also statistically significant with a small effect size ($\eta_p^2 = 0.033$). Figure 37 shows that the average of SE for $\hat{\theta}_{jM}$ is the smallest when the test contains 60% of double-coded items across different levels of testlet effect SD and different levels dual testlets correlation. The

variability of the averages of SE for $\hat{\theta}_{jM}$ at different levels of percentage of double-coded items is smaller when the testlet effect SD is smaller, meaning that the stability of $\hat{\theta}_{jM}$ across replications is more impacted by the percentage of double-coded items when the testlet effect variability is large. In addition, when testlet effect SD is small, the SE of $\hat{\theta}_{jM}$ is more impacted by the percentage of double-coded items as the dual testlet correlation increases— the $\hat{\theta}_{jM}$ is the most stable across replications while the testlet effect SD is small and the paired testlets are weakly correlated. When the testlet effect SD is large and the dual testlet correlation is 0.2, the $\hat{\theta}_{jM}$ are more stable when there are 20% or 40% double-coded items; when the dual testlet correlation increases, the $\hat{\theta}_{jM}$ starts to contain much more random error at 40% of double-coded items.

Table 17.
ANOVA Results of Significant Effects on the SE of $\hat{\theta}_{jM}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	104972.09	<0.001	0.854
model * testlet.sd	255.214	<0.001	0.014
model * dbcorr	382.93	<0.001	0.041
model * percent_dbcd	890.948	<0.001	0.09
model * testlet.sd * dbcorr	144.395	<0.001	0.016
model * testlet.sd * percent_dbcd	208.599	<0.001	0.023
model * dbcorr * percent_dbcd	465.741	<0.001	0.094
model * testlet.sd *dbcorr * percent_dbcd	181.004	<0.001	0.039
Between			
testlet.sd	949.02	<0.001	0.05
percent_dbcd	935.988	<0.001	0.094
dbcorr *percent_dbcd	145.417	<0.001	0.031
testlet.sd * dbcorr * percent_dbcd	154.178	<0.001	0.033

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

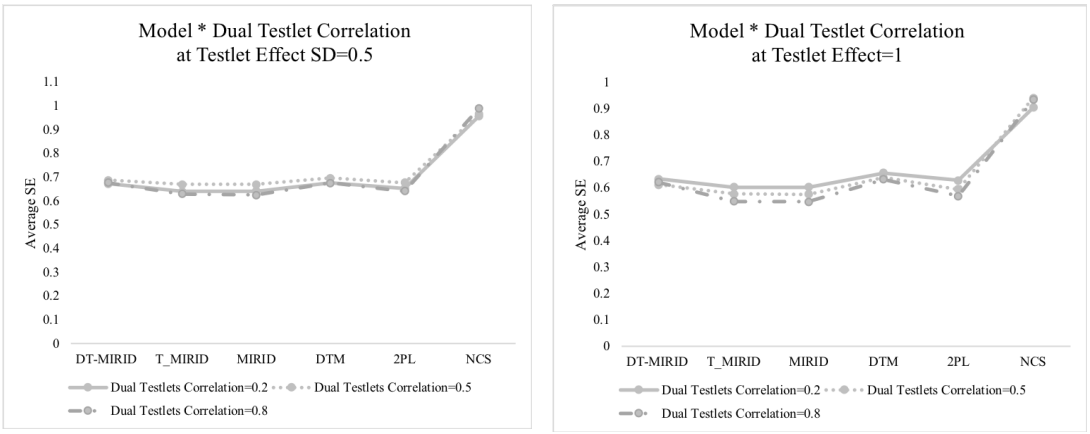


Figure 34. Significant three-way interaction effects on the SE of $\hat{\theta}_{jM}$—model*dual testlets correlation* testlet effect SD

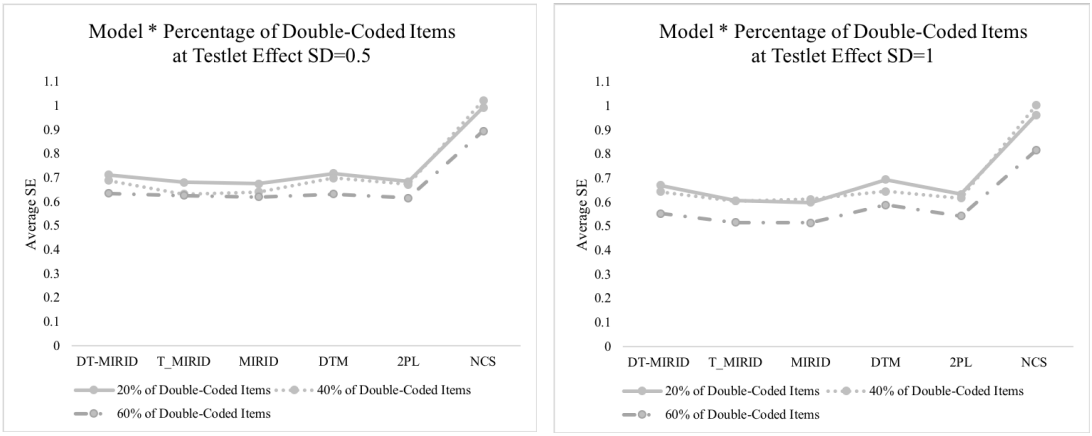


Figure 35. Significant three-way interaction effects on the SE of $\hat{\theta}_{jM}$—model*percentage of double-coded items* testlet effect SD

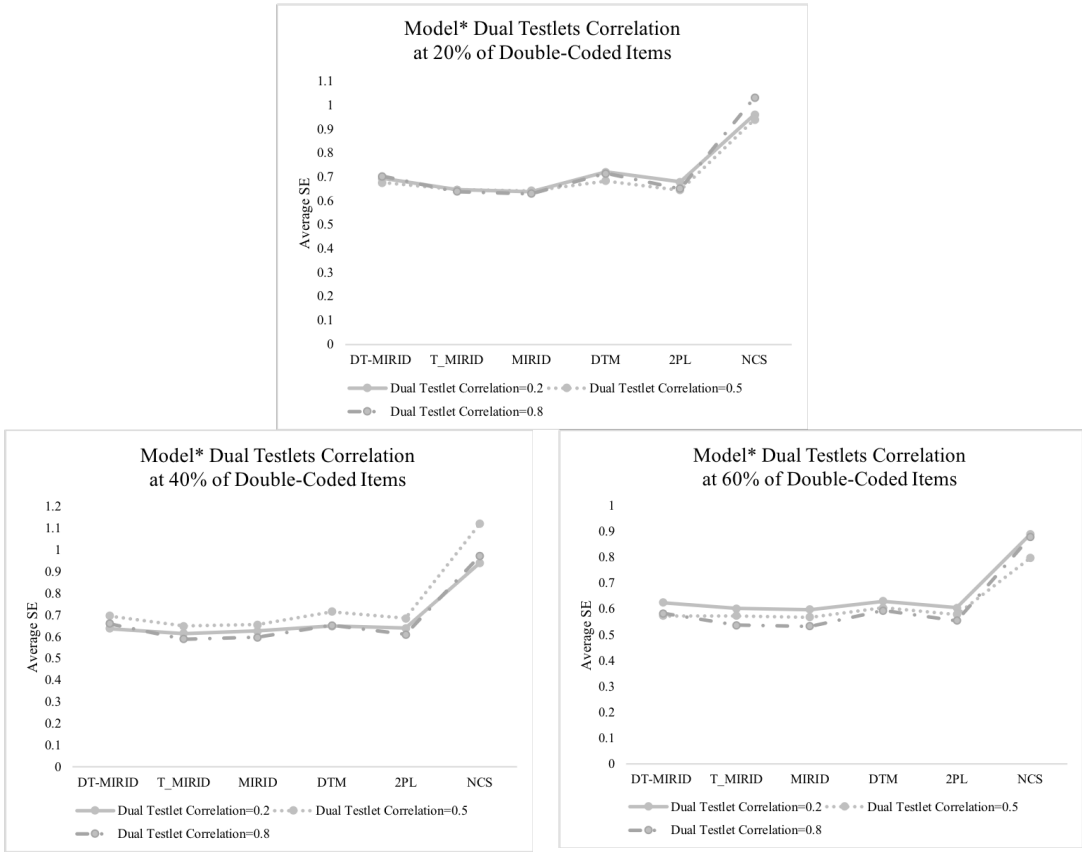


Figure 36. Significant three-way interaction effects on the SE of $\hat{\theta}_{jM}$— model*dual testlets correlation* percentage of double-coded items

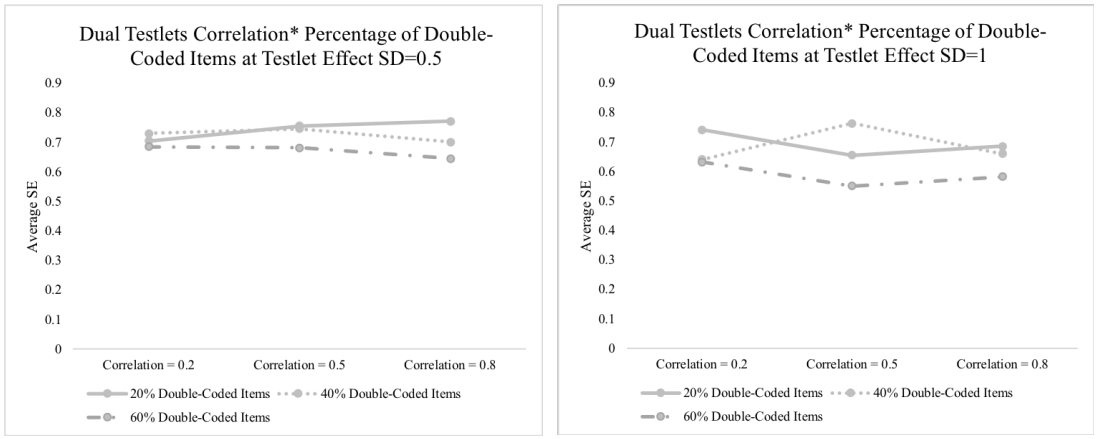


Figure 37. Significant three-way interaction effects on the SE of $\hat{\theta}_{jM}$—dual testlet correlation*percentage of double-coded items* testlet effect SD

In addition, the four two-way interactions have significant effects on the SE of $\hat{\theta}_{jM}$ — the interaction between model and testlet effect SD with a small effect size ($\eta_p^2 = 0.014$), the interaction between model and dual testlet correlation with a small

effect size ($\eta_p^2 = 0.041$), the interaction between model and the percentage of double-coded items with a medium effect size ($\eta_p^2 = 0.090$), and the interaction between dual testlet correlation and the percentage of double-coded items with a small effect size ($\eta_p^2 = 0.031$). Besides, model, testlet effect SD and the percentage of double-coded items are significant effects on the SE of $\hat{\theta}_{jM}$. A large effect is found for model on the SE of $\hat{\theta}_{jM}$ ($\eta_p^2 = 0.854$). The main effect of testlet effect SD ($\eta_p^2 = 0.050$) and the percentage of double-coded items ($\eta_p^2 = 0.094$) have small and medium effect size, repectively. The four-way interaction among model and all three manipulated factors is significant on the SE of $\hat{\theta}_{jM}$ with a small effect size ($\eta_p^2 = 0.039$).

RMSE. The ANOVA results are shown in Table 18. The three-way interaction among model, testlet effect SD and dual testlets correlation is a significant effect on the RMSE of $\hat{\theta}_{jM}$ with a small effect size ($\eta_p^2 = 0.010$). As Figure 38 presents, while the RMSEs of $\hat{\theta}_{jM}$ at different levels of dual testlet correlation become more similar for pattern scoring models while the testlet effect SD increases, NCS produces RMSEs of $\hat{\theta}_{jM}$ that are more diverged at levels of dual testlets correlation as the testlet effect variability increases. Moreover, the three-way interaction among model, dual testlets correlation and the percentage of double-coded items also significantly impacts the RMSEs of $\hat{\theta}_{jM}$ with a small effect size ($\eta_p^2 = 0.018$). Figure 39 shows that pattern scoring models yield RMSEs of $\hat{\theta}_{jM}$ that are less variable to the change in dual testlet correlation when a test contains 60% of double-coded items, while the variability of means of RMSEs for $\hat{\theta}_{jM}$ at levels of dual testlet correlation inflates when the percentage of double-coded items increases.

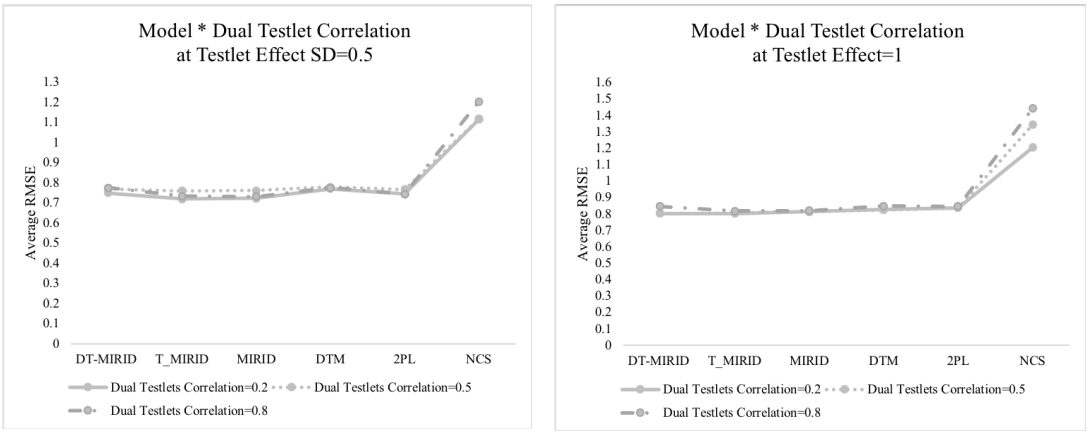


Figure 38. Significant three-way interaction effect on the RMSE of $\hat{\theta}_{jM}$— model* testlet effect SD* dual testlets correlation

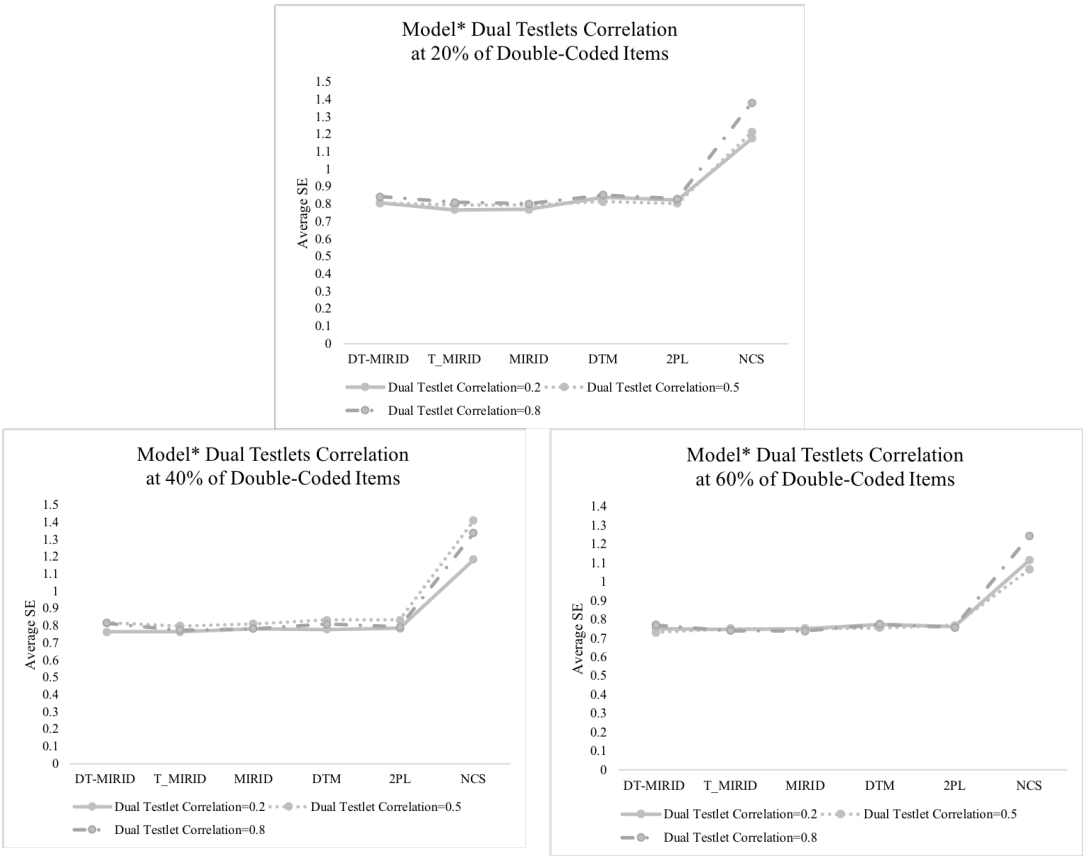


Figure 39. Significant three-way interaction effect on the RMSE of $\hat{\theta}_{jM}$— model* dual testlets correlation* percentage of double-coded items

In addition, the two-way interaction effects between model and testlet effect SD ($\eta_p^2 = 0.027$), between model and dual testlet correlation ($\eta_p^2 = 0.030$), and between model and percentage of double-coded items ($\eta_p^2 = 0.022$) are significant

with small effect sizes. The main effects of model, testlet effect SD and percentage of double-coded items are significant on the RMSE of $\hat{\theta}_{jM}$. The effect sizes for the model is large ($\eta_p^2 = 0.611$). The main effect of the testlet SD ($\eta_p^2 = 0.042$) and that of the percentage of double-coded items ($\eta_p^2 = 0.020$) have small effect sizes.

Table 18
ANOVA results of Significant Effects on the RMSE of $\hat{\theta}_{jM}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	28201.339	<0.001	0.611
model * testlet.sd	497.527	<0.001	0.027
model * dbcorr	273.606	<0.001	0.03
model * percent_dbcd	205.912	<0.001	0.022
model * testlet.sd * dbcorr	93.922	<0.001	0.01
model * dbcorr * percent_dbcd	81.319	<0.001	0.018
Between			
testlet.sd	795.869	<0.001	0.042
percent_dbcd	181.824	<0.001	0.02

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

In addition to the significant effects commonly observed in all subscores, a few unique significant effects for certain subscores are identified. The main effect of dual testlet correlation on the SE of $\hat{\theta}_{jD}$ is significant with a small effect size ($\eta_p^2 = 0.016$), the main effect of dual testlet correlation on the RMSE of $\hat{\theta}_{jS}$ is significant with a small effect size ($\eta_p^2 = 0.012$). The interaction effect between model and percentage of double-coded items on the bias of $\hat{\theta}_{jA}$ is statistically significant with a small effect size ($\eta_p^2 = 0.013$). The interaction between the testlet effect SD and the percentage of double-coded items has a small effect on the SE of $\hat{\theta}_{jA}$ ($\eta_p^2 = 0.010$) and the SE of $\hat{\theta}_{jS}$ ($\eta_p^2 = 0.013$). The four-way interaction effect on the RMSE of $\hat{\theta}_{jD}$

among the model and the three manipulated factors is significant with a small effect size ($\eta_p^2 = 0.019$).

Reliability

Overall Reliability

The results for the overall score reliability are summarized in Figure 40 (the overall score reliability and the subscore reliability for each study condition are tabulated in Appendix F). The overall reliability is calculated based on the overall ability parameters that are estimated using each of the comparison models.

Across all study conditions, NCS yields the highest reliability, followed by T-MIRID, MIRID and 2PL. NCS ability estimates are obtained by mapping the sum scores onto the TCC that is produced by integrating out the dual testlet effects. The T-MIRID ignores the dual testlet structure by assuming that the last 10 items belong to a third testlet and that the testlet effects from these three testlets are independent. The MIRID and 2PL ignore testlet structure by assuming LII, where a student's response to one item is not related to his/her response to another item after controlling for ability. Previous investigations have found that reliability is over estimated in the situation where LID is present but ignored (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Lukhele, 1997; Wainer & Thissen, 1996, Yen, 1993; Zenisky, Hambleton, & Sireci, 2002). Specifically, literature has suggested that ignoring item dependency inflates test information (Ip, 2000; Thissen, Steinberg, & Mooney, 1989; Reese, 1995; Wainer & Wang, 2001) and underestimates the SEM (Wainer, 1995; Wainer & Thissen, 1996). Results from the current study agree with previous studies.

Further, the current study shows that when testlet effects from different testlets correlate, ignoring the correlation between dual testlet effects will also result in spurious inflation of reliability estimates.

In addition, the reliability yielded from DTM decreases as the percentage of double-coded items increases. This is because DTM ignores the double-coded structure, therefore its ability estimates contain more error when there are more double-coded items. Whereas the reliability from the DTMIRID increases when the test contains more double-coded items. This phenomenon is consistent with what has been found in literature— large number of composite items (i.e. double-coded items in this study) will increase the accuracy of task weights estimates (Butter et al., 1998; Huang, 2011; Li, 2017). As a result, the estimation of ability parameters also contains less error. Even in MIRT framework, the use of double-coded items increases the score reliability estimates (Feinberg & Wainer, 2014).

Comparing the reliability produced by DT-MIRID and DTM, ignoring double-coded item structure reduces the overall score reliability, especially when there are more double-coded items in the test. The reliability of the overall ability estimates produced by the proposed model is higher when the dual testlets are less correlated and when the testlet effect SD is large.

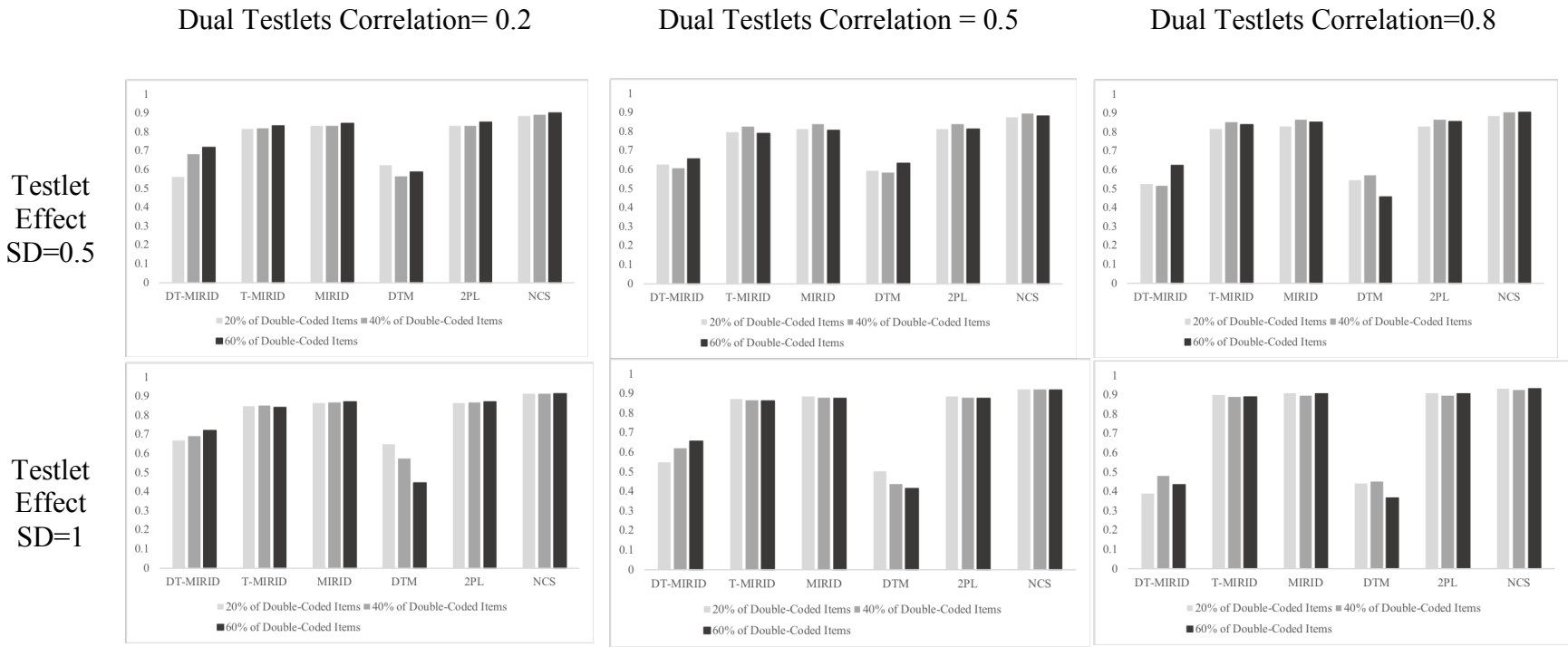


Figure 40. Reliability for overall score under each simulation condition for each model

Subscore Reliability

Since the subscore reliability is based on a subset of items from the test, the reliability for subscores is generally lower than the reliability of the overall score. The pattern found in subscore reliability is consistent across subscores. Therefore, this section only presents the reliability of subscore for subtraction as an example. Reliabilities for other subscores can be found in Appendix F.

Figure 41 presents the reliability for subscore of subtraction at each simulation condition for each model. Similar to patterns that have been identified in reliability for overall scores, the reliabilities produced by NCS, the T-MIRID, the MIRID and the 2PL are spuriously inflated. Reliabilities obtained from the proposed model are higher when the testlet effect SD is large and the correlation between dual testlet effects is small. That is, the subscore reliability tends to be large when testlet effects from dual testlets are less dependent, and the testlet effect variability is large. In addition, ignoring double-coded structure results in lower subscore reliability.

For subscore reliability, higher reliability tends to be obtained when there are 60% of double-coded items in the test. This pattern is true for all models, even for estimates yielded from DTM which should be penalized more when the percentage of double-coded items is high, in some conditions (e.g. testlet effect SD=0.5, dual testlets correlation=0.2). Although this pattern is also found in the overall score reliability based on estimates yielded from the DT-MIRID, the two may be due to different reasons—the subscore reliability is estimated with more items, if the percentage of double-coded items is high, whereas the test length is constant for all

levels of percentage of double-coded items in estimating overall score reliability. Take subscore of subtraction as an example. (Information for items included in subscore of subtraction is presented in Table 19.) There are 12 items for the subscore of subtraction when there are 40% and 60% of double-coded items, yet only 9 items when there are 20% of double-coded items. Item information for other subscores can be found in Appendix G. This means that when investigating the impact of the percentage of double-coded items on the subscore reliability, there are two impacts that are inseparable—(a) the impact of the number of items in the subscore, and (b) the impact of the parameter estimation for the component difficulty, task weight and intercept for the composite item.

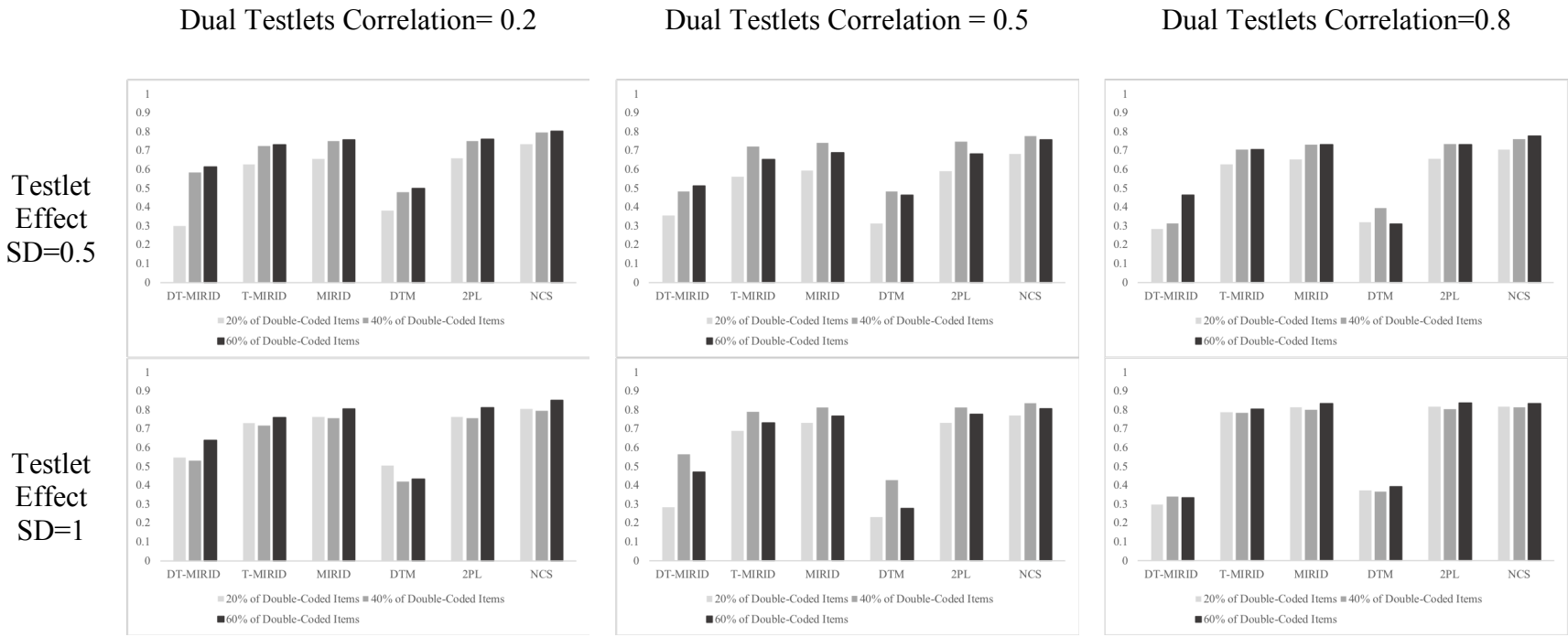


Figure 41. Subscore reliability for subscore of subtraction under each simulation condition for each model

Table 19

Information for Items in Subscore of Subtraction

	20% of Double-coded Items			40% of Double-coded Items			60% of Double-coded Items		
Position in subscore	Position in Test	Composite Item	Item Difficulty Used	Position in Test	Composite Item	Item difficulty Used	Position in Test	Composite Item	Item Difficult Used
1	3	no	3	3	no	3	2	no	2
2	4	no	4	4	no	4	5	yes	2
3	9	yes	3	7	yes	3	8	yes	2
4	13	no	13	8	yes	4	9	yes	2
5	14	no	14	13	no	13	12	no	12
6	20	yes	13	14	no	14	15	yes	12
7	23	no	23	19	yes	13	18	yes	12
8	24	no	24	20	yes	14	19	yes	12
9	29	yes	23	21	no	21	22	no	22
10				22	no	22	25	yes	22
11				27	yes	21	28	yes	22
12				28	yes	22	29	yes	22

Note: 1. For component items, numbers in "Item Difficulty Used" are positions of the items in the test. For a composite item, the number in column 4, 7 and 10 is the position of the component item that assesses subtraction.

2. The scale of shades indicates which testlet the item belongs to, under the true condition. The lightest shade indicates the first testlet, the medium shade indicates the second testlet and the darkest indicates items for the dual testlets.

Model Fit

As described in Chapter 3, the proportion of identifying the true model as the best fitting model using AIC, BIC and DIC under each study condition is used as criterion to evaluate the model fit indices. However, none of the three indices identifies the true model as the best fitting model for any replication in 16 out of 18 study conditions. Therefore, the proportion of identifying each model as the best fitting model under each condition is summarized in this section. Table 20 presents the percentage of identifying each model as the best fitting model for all study conditions.

The AIC and BIC have identified same model as the best fitting model in all study conditions. The AIC and BIC consistently identify the T-MIRID as the best fitting model for almost all study conditions, except that they choose DTM as the best fitting model in the conditions where the testlet effect SD is 0.5 and there are 60% of double-coded items and the dual testlets correlation is 0.2 or 0.5. The DIC favors the 2PL, followed by the MIRID for most study conditions. The proposed model was selected as the best fitting model only by DIC and only in the conditions with the testlet effect SD being 1, the dual testlets correlation being 0.2, and there being 20% or 60% of double-coded items.

The model index was calculated based on the item and overall ability parameter estimation; it does not assess the data-model fit in estimating subscores. In estimating model parameters, there are two pairs of models that are theoretically very similar in terms of overall fit—(a) DTMIRID and DTM, and (b) MIRID and 2PL.

The difference between models within each pair is that the simpler model does not decompose composite item difficulty, while the more complex model does. Such difference has little impact on the data-model fit, but the more complex model is penalized more as it contains more parameters. That is why the DIC struggles (judging from the percentages of being identified as the best fitting model) between MIRID and 2PL and between the DT-MIRID and the DTM in identifying the best fitting models. However, these extra parameters in the more complex model are necessary in subscore reporting. Hence, model fit index cannot be used as the sole criterion in model selection.

In summary, the AIC and BIC fail to identify the proposed model as the best fitting model for all study conditions, and DIC is only able to identify the proposed model as the best fitting model when the testlet effect variability is large and the dual testlets correlation is small. This indicates that these model fit indices are limited in empirical evaluation of data-model fit for the proposed model. In addition, since one of the main purpose of the proposed model is to report subscores, sacrificing model fit by including the necessary parameters for subscore estimation has many gains. In other words, model fit indices should not be emphasized in evaluating model performance, especially in terms of subscore reporting for the proposed method.

Table 20

Proportion of Identifying Each Model as the Best Fitting Model

Conditions			AIC					BIC					DIC				
VAR_γ	$\rho_{\gamma_1\gamma_2}$	% Double-coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	DT-MIRID	T-MIRID	MIRID	DTM	2PL	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0	1	0	0	0	0	1	0	0	0	0	0	0.43	0	0.57
		40%	0	1	0	0	0	0	1	0	0	0	0	0	0.67	0	0.33
		60%	0	0.03	0	0.97	0	0	0.03	0	0.97	0	0	0	0	0	1
	0.5	20%	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
		40%	0	1	0	0	0	0	1	0	0	0	0	0	0.43	0	0.57
		60%	0	0.17	0	0.83	0	0	0.17	0	0.83	0	0	0	0	0	1
	0.8	20%	0	1	0	0	0	0	1	0	0	0	0	0	0.5	0	0.50
		40%	0	1	0	0	0	0	1	0	0	0	0	0	0.53	0	0.47
		60%	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
1	0.2	20%	0	1	0	0	0	0	1	0	0	0	0.60	0	0	0.40	0
		40%	0	1	0	0	0	0	1	0	0	0	0.43	0	0	0.57	0
		60%	0.13	0.80	0	0.07	0	0.13	0.80	0	0.07	0	0.53	0	0	0.43	0.03
	0.5	20%	0	1	0	0	0	0	1	0	0	0	0	0	0.63	0	0.37
		40%	0	1	0	0	0	0	1	0	0	0	0.07	0	0.03	0.03	0.87
		60%	0.07	0.90	0	0.03	0	0.07	0.90	0	0.03	0	0	0	0.03	0.07	0.90
	0.8	20%	0	1	0	0	0	0	1	0	0	0	0	0	0.50	0	0.50
		40%	0	1	0	0	0	0	1	0	0	0	0	0	0.23	0	0.77
		60%	0	0.63	0	0.37	0	0	0.67	0	0.33	0	0	0	0	0	1

Note: The grey shading indicates the highest proportion identified among models by each model fit index under each study condition.

Chapter 5: Discussion

This study proposes the 2PL-DT-MIRID to model complex testlet structure where correct item responses require information from paired testlets and to report content-based subscores by decomposing composite item difficulty into content-specific component difficulties. A simulation study is conducted to investigate the performance of the proposed model in comparison with other underspecified models. In addition, overall and subdomain abilities are also estimated using NCS. Overall number-correct scores and subscores are compared with those from pattern scoring. The impact of testlet effect SD, dual testlets correlation and the percentage of double-coded items are evaluated together with the impact from mis-specification in model structure. The model performance is assessed by parameter estimation accuracy, score reliability and model selection. The findings from this study are compared with findings from previous investigations and summarized in a systematic way in the hope that this study will serve as a reference for future exploration.

The Simulation Results

Based on the results presented in Chapter 4, this section summarizes the findings from four perspectives— (a) the impact of ignoring dual testlet or/and double-coded items on model parameter recovery and subscore estimation, (b) the impact of the manipulated factors on item and overall ability parameter recovery and subscore estimation, (c) the implications on the score reliabilities, and (d) the implications on the effectiveness of different model selection indices.

Impact of Ignoring Dual Testlets or/and Double-Coded Items

The data generating model outperforms other underspecified models (i.e. the T-MIRID, the MIRID, the DTM and the 2PL) in terms of model parameter estimation. The proposed model has, on average, the smallest bias, SE and RMSE for item discrimination parameters, the smallest bias and RMSE for item difficulty parameters, and the smallest RMSE for task weights, tau, correlation between testlet effects for the paired testlets. The proposed model and the DTM performed equally well on overall ability parameter recovery. Although the DTM produces the smallest SE and RMSE on average, the difference between the average SE and RMSE produced by the DTM and those by the proposed model is to the third decimal places and the direction of the difference varies for different study conditions.

In subscore estimation, however, the T-MIRID and the MIRID perform the best, followed by the proposed DT-MIRID. The performance of the DTM and the 2PL is much worse than models that properly accommodate double-coded items (i.e. T-MIRID, MIRID and DT-MIRID). The NCS is the worst in terms of score estimation among the six methods by yielding the largest bias, SE and RMSE.

Based on the study results, ignoring the dual testlet structure has a major impact on item parameter recovery. Previous studies have concluded that not modeling the dependency among items due to testlets while it exists results in underestimated item discrimination (Bradlow et al., 1999; Wainer et al., 2000) and shrinking variance of item difficulties (Ackerman, 1987; Bradlow et al., 1999; Reese, 1995). Jiao et al. (2017) found that ignoring dual testlets will result in negatively biased $\hat{a}_i$ averaging across all items, but they also found that $\hat{a}_i$ and $\hat{b}_i$ for items that

require information from both testlets are over estimated in much larger magnitude when the dual testlets are not accommodated, as comparing to the bias of $\hat{a}_i$ and $\hat{b}_i$ for items nested within a single testlet. The current study observes the same pattern, where $\hat{a}_i$ and $\hat{b}_i$ for items that require information from dual testlets are estimated with larger positive bias and larger RMSEs than those for items in a single testlet. Since one-third of the items in the current study are based on dual testlets whereas only one-seventh of the total items in Jiao et al. (2017) are based on paired-testlets, the average bias for $\hat{a}_i$ in the current study is influenced more by the large positive bias found in items based on dual testlets when ignoring dual testlets for all study conditions. In addition, ignoring dual testlet structure also leads to more error in the estimation of task weights and the intercepts.

The impact of ignoring testlet effects from dual testlets on the overall ability parameter estimation is also different from the impact of ignoring testlet effects that are not correlated between testlets. The literature suggests that the ability estimates are more spread-out than the true parameters (Ackerman, 1987; Reese, 1995). In our study, the theta distributions are attenuated towards the mean for all models across all study conditions. When ignoring the dual testlets, the ability estimates are less attenuated to the mean than ability estimates yielded from the DT-MIRID and the DTM do.

Ignoring double-coded items has a major impact on the estimation of subscores. Such impact is consistent on all four subscores and it is entirely anticipated. The double-coded items are designed to assess two arithmetic operations in one item. The DTM and the 2PL will count the double-coded item twice, once for

estimating each subscore. Therefore, errors are introduced by including students' ability for the off-target subdomain in the on-target subscore. Although the NCS subscores contain even more error, the errors contained in the NCS subscores are mainly due to ignoring response pattern; in other words, the NCS weights each item equally in estimating the logit scores when some items should contribute more to students' scores.

Impact of Manipulated Factors

The testlet effect SD has significant impact on almost all model parameters and all subscores. Judging from the SE and the RMSE, the larger the testlet effect SD, the more accurate the item discrimination estimates, the testlet variance estimates and the dual testlets correlation estimates. Bradlow et al. (1999) found that the impact of ignoring testlet effects for a single testlet (as opposed to dual testlets where testlet effects correlate) is larger on item parameters, when the testlet effect variability increases. The current study also finds that the impact of ignoring dual testlets on item discrimination and testlet effect variance is more severe when the testlet SD is larger. In terms of the overall ability and subscores, smaller testlet effect SD results in less total error in estimation.

Jiao et al. (2017) found that a smaller correlation between testlet effects from the dual testlets associates with larger bias in testlet variance estimates. However, the average bias in Jiao et al. (2017) was calculated across the variances across all testlets including independent testlets and the two testlets that are correlated. When only looking at the two testlets that are correlated, as in the current study, their findings agree with the current study—smaller correlation between dual testlets leads to

smaller positive bias. Nevertheless, different from Jiao et al. (2017), the current study finds that larger dual testlet correlation improves the recovery of the correlation between testlet effects from the dual testlets, judging from the bias, SE and RMSE. For the overall ability parameters, less error is obtained when the dual testlets are less correlated. On the items side, the item discrimination estimates, task weights and the intercept obtained in conditions where the dual testlets are less correlated contain less error, according to RMSE. In practice, high correlation between dual testlets should be rare. Since the goal of adopting dual testlets in a test is to assess students' ability in synthesizing information from different sources, if the dual testlets provide similar information, there is little gain in using dual testlets in a test.

The percentage of double-coded items has little impact on the model parameter estimation, but it influences the subscore estimation. According to the bias, SE and RMSE, large percentage of double-coded items increases the accuracy of subscore estimation. The reason is that having more double-coded items in the tests means that each component item difficulty is used more frequently in double-coded items. This is essentially to test and to retest students' ability on the same knowledge/skill. Therefore, the estimated subscores are more accurate. Although including more double-coded items can improve subscore estimation accuracy, having too many double-coded items is neither efficient in content coverage nor economical in developing high quality items. To decide on the percentage of double-coded items, test developers should balance among adequacy in content coverage, assessment on integrated skills, and accuracy in reported subscores.

Score Reliability

Based on the results, three major findings are summarized on score reliability. First, ignoring dual testlets naively inflates the reliability as the correlation between item responses and the ability estimates increases due to the failure in separating the ability from the testlet effect. Second, ignoring the double-coded items will negatively impact the score reliability based on the model parameter estimates, because subdomain abilities are contaminated by information from other domains. And third, high percentage of double-coded items increases the score reliability by providing consistent subdomain estimates with more frequent use of component item difficulties.

As previously discussed, the variance of the overall ability estimates shrinks for the proposed model. One caveat in calculating score reliability is that the attenuation of the ability towards the mean reduces the reliability estimates. This is why the reliability of the NCS scores with a variance closer to 1 are much higher than that calculated by using the DT-MIRID, even though they use the same estimated item parameters.

Model Selection

This study evaluates model fit using three commonly used model fit indices—AIC, BIC and DIC. Result shows that AIC and BIC fail to identify the proposed model as the best fitting model in all study conditions, and the DIC only identifies the proposed model as the best fitting model when the testlet effect SD is 1, the dual testlet correlation is 0.2 and there are 20% or 60% of double-coded items in the test. On one hand, these three indices are ineffective in identifying the model fit

improvement from ignoring dual testlet effects to accommodating the dual testlets and from ignoring double-coded items to modeling the double-coded items explicitly in the model. On the other hand, since decomposing item difficulty is necessary for subscore reporting, the fact that the commonly used fit index in Bayesian estimation cannot identify the true model may not jeopardize the utility of the proposed model. A new model fit index that assesses the overall model fit including the fit in model parameter estimation and the fit in subscore estimation can be developed for reporting subscores with the proposed model. In addition, other model fit indices can be compared in a more comprehensive study for investigating model selection for the proposed method.

Limitations and Future Investigations

Like every study, certain limitations remain in the current study. This study simulates response data for a test containing two testlets, each with 10 items and an additional set of items, 21-30 requiring information from both testlets. Although such a design was adopted for a focused investigation on the dual testlets structure and the double-coded items, it is not the most realistic in practice. In a real testing scenario, a test often contains single-coded items that are not nested within testlets and testlets that do not correlate with other testlets. The average error measures reported in the current study magnified the impact of double-coded items and the dual testlet comparing to what would have presented in a real testing scenario. Future studies may consider including single-coded items that do not belong to any testlets and independent testlets to the test structure to assess the impact of double-coded items

and that of the dual testlets from a more realistic perspective. Results based on a more realistic test design may have more immediate implication to test operations.

The current study only investigated the impact of testlet effect SD, dual testlet correlation and the percentage of double-coded items. Such exploration is far from enough for validating the use of such model in a large-scale test. Factors such as sample size, test length and number of items in a testlet could also be added in future exploration.

In addition to limitations in the study design, a caveat in the method should also be mentioned. From calculating the empirical true subscores to estimate model parameters and to subscore estimation, the complexity in the procedures may introduce more random error in the final subscore estimates. Future investigation could compare subscores produced using procedures in the current study with subscores yielded using other methods, such as MIRT to make relative statement of the subscore estimation accuracy.

Although there are limitations, the contributions of this study are notable. This study is motivated by innovative item types in the test where a single test item contributes to two subscores and the dual testlets are embedded in the test to assess students' ability on information synthesis. The DT-MIRID proposed in this study explicitly accommodates the dual testlet structure and the double-coded items in the model structure for providing accurate estimates on the overall score and subscores. The DT-MIRID estimates subscores from a new perspective— decomposing item difficulty parameters into component difficulties that are content domain specific. Different from MIRT, the proposed method is more versatile in that it can be used to

produce both pattern scores and number-correct scores. Consequence of ignoring complex testlet and item structures are modeled and assessed under study conditions that vary in terms of testlet effect SD, dual testlets correlation and percentage of double-coded items. Results of this study show that parameter estimation accuracy for item parameters, overall ability parameters, and subscores are all improved by accommodating complex testlet and item structure in innovative item types.

Appendix A Data Generating Models for the Test with 20% of Double-Coded Items

Item	Testlet	Arithmetic Operation(s)	Item Type	Data Generating Models
1	1	A	Component, Single Testlet	$\mathrm{P}(\mathrm{X_{jik}}=1)=\frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}$
2	1	A	Component, Single Testlet	Same as item 1
3	1	S	Component, Single Testlet	Same as item 1
4	1	S	Standalone, Single Testlet	$\mathrm{P}(\mathrm{X_{ji0}}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{i1}\right)\right)}$
5	1	M	Component, Single Testlet	Same as item 1
6	1	M	Standalone, Single Testlet	Same as item 4
7	1	D	Standalone, Single Testlet	Same as item 4
8	1	D	Standalone, Single Testlet	Same as item 4
9	1	A (1) & S (3)	Double-Coded Single Testlet	$\mathrm{P}(\mathrm{X_{ji0}}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}$
10	1	A (2) & M (5)	Double-Coded Single Testlet	Same as in item 9
11	2	A	Component, Single Testlet	$\mathrm{P}(\mathrm{X_{jik}}=1)=\frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{2(i)}}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{2(i)}}-\beta_{ik}\right)\right)}$
12	2	A	Standalone, Single Testlet	$\mathrm{P}(\mathrm{X_{ji0}}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\beta_{i1}\right)\right)}$
13	2	S	Component, Single Testlet	Same as item 11

14	2	S	Standalone, Single Testlet	Same as item 12
15	2	M	Component, Single Testlet	Same as item 11
16	2	M	Standalone, Single Testlet	Same as item 12
17	2	D	Component, Single Testlet	Same as item 11
18	2	D	Standalone, Single Testlet	Same as item 12
19	2	A (11) & D (17)	Double-Coded Single Testlet	$P(X_{ji0}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}$
20	2	S (13) & M (15)	Double-Coded Single Testlet	Same as item 19
21	1 & 2	A	Standalone, Paired Testlet	$P(X_{ji0}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{i1}\right)\right)}*\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\beta_{i1}\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2(i)}}-\beta_{i1}\right)\right)}$
22	1 & 2	A	Standalone, Paired Testlet	Same as item 21
23	1 & 2	S	Component, Paired Testlet	$P(X_{jik}=1)=\frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{1(i)}}-\beta_{ik}\right)\right)}*\frac{\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{2}(i)}-\beta_{ik}\right)\right)}{1+\exp\left(a_{ik}\left(\theta_j+\gamma_{jd_{2}(i)}-\beta_{ik}\right)\right)}$
24	1 & 2	S	Standalone, Paired Testlet	Same as item 21
25	1 & 2	M	Component, Paired Testlet	Same as item23
26	1 & 2	M	Standalone, Paired Testlet	Same as item 21
27	1 & 2	D	Component, Paired Testlet	Same as item23
28	1 & 2	D	Component, Paired Testlet	Same as item23

29	1 & 2	S (23) & D (27)	Double-Coded, Paired Testlet	$P(X_{ji0}=1)=\frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{1(i)}}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)} * \frac{\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2}(i)}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}{1+\exp\left(a_{i0}\left(\theta_j+\gamma_{jd_{2}(i)}-\sum_{k=1}^{K}\sigma_k\beta_{ik}-\tau\right)\right)}$
30	1 & 2	M (25) & D (28)	Double-Coded, Paired Testlet	Same as item 29

Appendix B Derivation of Item Information for 2PL-DT-MIRID

For 2PL-DT- MIRID, Define

$$\begin{cases} exp1 = \exp\left(a_{i0}(\theta_j + \gamma_{jd_{1(i)}} - \sum_{k=1}^{k} \sigma_k \beta_{ik} - \tau)\right) \\ exp1 = \exp\left(a_{i0}(\theta_j + \gamma_{jd_{2(i)}} - \sum_{k=1}^{k} \sigma_k \beta_{ik} - \tau)\right) \end{cases},$$

then

$$\begin{cases} P(\theta_j) = \dfrac{\exp_1}{(1+\exp_1)} \dfrac{\exp_2}{(1+\exp_2)} \\ Q(\theta_j) = 1 - P(\theta_j) = \dfrac{1+\exp_1+\exp_2}{(1+\exp_1)(1+\exp_2)} \end{cases}.$$

The definition of item information is

$$I(\theta_j) = \frac{\left(P'(\theta_j)\right)^2}{P(\theta_j)Q(\theta_j)}.$$

In the case of 2PL-DT-MIRID,

$$\begin{aligned} & P'(\theta_j) \\ &= \frac{-\exp_1\exp_2\left((1+\exp_1)(1+\exp_2)\right)' + (\exp_1\exp_2)'\left((1+\exp_1)(1+\exp_2)\right)}{\left((1+\exp_1)(1+\exp_2)\right)^2} \\ &= \frac{\exp_1\exp_2' + \exp_1{}'\exp_2 + \exp_1'\exp_2^2 + \exp_1^2\exp_2{}'}{\left((1+\exp_1)(1+\exp_2)\right)^2} \\ &= \frac{a_{i0}\exp_1\exp_2(2+\exp_1+\exp_2)}{\left((1+\exp_1)(1+\exp_2)\right)^2} \\ &= \frac{a_{i0}\exp_1\exp_2\left(\frac{1}{1+\exp_1}+\frac{1}{1+\exp_2}\right)}{(1+\exp_1)(1+\exp_2)} \end{aligned}.$$

Therefore,

$$I(\theta_j) = \frac{\left(P'(\theta_j)\right)^2}{P(\theta_j)Q(\theta_j)}$$

$$= \frac{\dfrac{a_{i0}^2 \exp_1^2 \exp_2^2 \left(\dfrac{1}{1+\exp_1} + \dfrac{1}{1+\exp_2}\right)^2}{(1+\exp_1)^2(1+\exp_2)^2}}{\dfrac{\exp_1 \exp_2 (1+\exp_1+\exp_2)}{(1+\exp_1)^2(1+\exp_2)^2}}$$

$$= \frac{a_{i0}^2 \exp_1 \exp_2 \left(\dfrac{1}{1+\exp_1} + \dfrac{1}{1+\exp_2}\right)^2}{1+\exp_1+\exp_2}$$

Appendix C Bias, SE and RMSE for Each Model Parameter and Subscores

Table 21

Bias for Item Discrimination Parameter under Each Study Condition

Manipulated Factors			Bias				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	-0.177	0.157	0.191	-0.140	0.199
		40%	-0.098	0.134	0.167	-0.169	0.163
		60%	-0.096	0.128	0.160	-0.174	0.168
	0.5	20%	-0.081	0.146	0.178	-0.098	0.185
		40%	-0.147	0.160	0.198	-0.160	0.198
		60%	-0.088	0.136	0.174	-0.084	0.200
	0.8	20%	-0.150	0.170	0.212	-0.141	0.215
		40%	-0.223	0.171	0.218	-0.195	0.226
		60%	-0.145	0.137	0.183	-0.223	0.181
1	0.2	20%	-0.022	0.278	0.267	-0.036	0.266
		40%	0.000	0.265	0.259	-0.080	0.253
		60%	0.031	0.289	0.313	-0.130	0.309
	0.5	20%	-0.085	0.361	0.378	-0.109	0.375
		40%	-0.026	0.343	0.357	-0.121	0.351
		60%	0.007	0.375	0.397	-0.125	0.397
	0.8	20%	-0.169	0.455	0.489	-0.146	0.490
		40%	-0.085	0.417	0.449	-0.098	0.443
		60%	-0.148	0.483	0.529	-0.168	0.534

Table 22

SE for Item Discrimination Parameter under Each Study Condition

Manipulated Factors			SE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.094	0.125	0.124	0.098	0.128
		40%	0.088	0.104	0.104	0.081	0.106
		60%	0.111	0.099	0.099	0.084	0.107
	0.5	20%	0.085	0.108	0.107	0.085	0.111
		40%	0.082	0.107	0.107	0.082	0.113
		60%	0.081	0.102	0.102	0.088	0.114
	0.8	20%	0.078	0.107	0.107	0.080	0.111
		40%	0.080	0.111	0.112	0.086	0.124
		60%	0.078	0.101	0.102	0.075	0.110
1	0.2	20%	0.099	0.119	0.110	0.096	0.112
		40%	0.094	0.112	0.106	0.094	0.108
		60%	0.091	0.098	0.095	0.080	0.102
	0.5	20%	0.088	0.112	0.112	0.087	0.113
		40%	0.090	0.113	0.109	0.086	0.112
		60%	0.099	0.117	0.113	0.099	0.120
	0.8	20%	0.084	0.123	0.119	0.084	0.123
		40%	0.084	0.119	0.114	0.083	0.117
		60%	0.109	0.126	0.125	0.085	0.134

Table 23

RMSE for Item Discrimination Parameter under Each Study Condition

Manipulated Factors			RMSE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.206	0.243	0.260	0.179	0.269
		40%	0.135	0.202	0.220	0.189	0.218
		60%	0.159	0.234	0.251	0.195	0.217
	0.5	20%	0.123	0.223	0.241	0.134	0.229
		40%	0.169	0.214	0.238	0.181	0.240
		60%	0.138	0.240	0.254	0.123	0.245
	0.8	20%	0.170	0.219	0.249	0.163	0.253
		40%	0.237	0.223	0.256	0.215	0.268
		60%	0.166	0.194	0.224	0.235	0.223
1	0.2	20%	0.102	0.309	0.309	0.104	0.308
		40%	0.097	0.293	0.286	0.126	0.282
		60%	0.099	0.312	0.334	0.154	0.331
	0.5	20%	0.124	0.381	0.398	0.141	0.395
		40%	0.095	0.365	0.377	0.149	0.371
		60%	0.102	0.397	0.417	0.161	0.418
	0.8	20%	0.190	0.473	0.505	0.169	0.506
		40%	0.120	0.435	0.464	0.132	0.459
		60%	0.190	0.514	0.554	0.192	0.552

Table 24

Bias for Item Difficulty Parameter under Each Study Condition

Manipulated Factors			Bias				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.011	0.288	0.276	0.007	0.277
		40%	-0.003	0.313	0.302	-0.007	0.293
		60%	-0.077	0.245	0.236	-0.009	0.245
	0.5	20%	-0.007	0.218	0.206	-0.007	0.256
		40%	-0.001	0.304	0.289	-0.013	0.283
		60%	-0.073	0.197	0.186	-0.009	0.197
	0.8	20%	-0.006	0.302	0.286	-0.008	0.295
		40%	0.000	0.244	0.231	-0.019	0.232
		60%	-0.026	0.340	0.322	-0.016	0.340
1	0.2	20%	0.004	0.277	0.269	0.005	0.273
		40%	0.003	0.299	0.288	-0.002	0.311
		60%	0.010	0.225	0.205	0.003	0.225
	0.5	20%	-0.005	0.259	0.249	-0.011	0.249
		40%	-0.003	0.277	0.265	-0.013	0.251
		60%	-0.011	0.193	0.190	-0.025	0.193
	0.8	20%	0.005	0.248	0.236	-0.002	0.239
		40%	-0.012	0.227	0.216	-0.013	0.240
		60%	0.049	0.245	0.231	-0.035	0.245

Table 25
SE for Item Difficulty Parameter under Each Study Condition

Manipulated Factors			SE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.106	0.102	0.097	0.106	0.099
		40%	0.105	0.097	0.095	0.117	0.098
		60%	0.138	0.078	0.076	0.113	0.078
	0.5	20%	0.108	0.103	0.098	0.114	0.097
		40%	0.099	0.088	0.084	0.105	0.094
		60%	0.109	0.089	0.084	0.115	0.089
	0.8	20%	0.114	0.112	0.105	0.116	0.108
		40%	0.093	0.076	0.072	0.100	0.079
		60%	0.105	0.082	0.078	0.115	0.082
1	0.2	20%	0.098	0.080	0.076	0.101	0.078
		40%	0.102	0.087	0.082	0.113	0.085
		60%	0.094	0.073	0.068	0.121	0.073
	0.5	20%	0.105	0.078	0.074	0.108	0.077
		40%	0.092	0.072	0.068	0.105	0.072
		60%	0.084	0.073	0.070	0.113	0.073
	0.8	20%	0.103	0.064	0.061	0.101	0.064
		40%	0.090	0.058	0.055	0.096	0.057
		60%	0.124	0.053	0.050	0.096	0.053

Table 26

RMSE for Item Difficulty Parameter under Each Study Condition

Manipulated Factors			RMSE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.201	0.358	0.334	0.168	0.335
		40%	0.131	0.379	0.368	0.189	0.364
		60%	0.177	0.318	0.308	0.217	0.318
	0.5	20%	0.154	0.352	0.342	0.173	0.335
		40%	0.161	0.366	0.346	0.172	0.353
		60%	0.198	0.399	0.388	0.162	0.399
	0.8	20%	0.199	0.379	0.356	0.189	0.364
		40%	0.244	0.312	0.309	0.218	0.308
		60%	0.190	0.424	0.433	0.270	0.424
1	0.2	20%	0.107	0.354	0.367	0.115	0.373
		40%	0.106	0.412	0.422	0.152	0.431
		60%	0.098	0.396	0.429	0.186	0.396
	0.5	20%	0.138	0.373	0.383	0.155	0.383
		40%	0.101	0.398	0.413	0.167	0.392
		60%	0.090	0.264	0.261	0.185	0.264
	0.8	20%	0.206	0.369	0.378	0.185	0.383
		40%	0.145	0.379	0.394	0.154	0.410
		60%	0.237	0.336	0.344	0.183	0.336

Table 27
Bias for Task weight under Each Study Condition

Manipulated Factors			Bias				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	-0.026	-0.093	-0.095	NA	NA
		40%	0.018	0.072	0.072	NA	NA
		60%	-0.043	0.004	0.005	NA	NA
	0.5	20%	0.031	-0.033	-0.030	NA	NA
		40%	-0.018	-0.042	-0.044	NA	NA
		60%	-0.085	-0.040	-0.044	NA	NA
	0.8	20%	0.027	-0.002	-0.006	NA	NA
		40%	0.006	0.043	0.040	NA	NA
		60%	-0.031	0.112	0.110	NA	NA
1	0.2	20%	-0.012	-0.027	-0.024	NA	NA
		40%	0.006	-0.015	-0.021	NA	NA
		60%	0.000	0.004	0.011	NA	NA
	0.5	20%	0.015	0.025	0.027	NA	NA
		40%	0.009	0.157	0.162	NA	NA
		60%	-0.013	0.014	0.014	NA	NA
	0.8	20%	0.020	-0.007	-0.007	NA	NA
		40%	-0.023	-0.055	-0.057	NA	NA
		60%	-0.128	-0.210	-0.214	NA	NA

Table 28
SE for Task weight under Each Study Condition

Manipulated Factors			SE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.082	0.083	0.083	NA	NA
		40%	0.071	0.065	0.065	NA	NA
		60%	0.106	0.063	0.065	NA	NA
	0.5	20%	0.053	0.042	0.043	NA	NA
		40%	0.055	0.065	0.064	NA	NA
		60%	0.045	0.059	0.057	NA	NA
	0.8	20%	0.089	0.083	0.082	NA	NA
		40%	0.050	0.058	0.058	NA	NA
		60%	0.050	0.052	0.052	NA	NA
1	0.2	20%	0.071	0.070	0.070	NA	NA
		40%	0.039	0.041	0.043	NA	NA
		60%	0.042	0.031	0.031	NA	NA
	0.5	20%	0.098	0.090	0.091	NA	NA
		40%	0.046	0.055	0.055	NA	NA
		60%	0.037	0.034	0.035	NA	NA
	0.8	20%	0.085	0.074	0.073	NA	NA
		40%	0.053	0.067	0.068	NA	NA
		60%	0.060	0.050	0.049	NA	NA

Table 29
RMSE for Task weight under Each Study Condition

Manipulated Factors			RMSE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.115	0.143	0.144	NA	NA
		40%	0.074	0.113	0.113	NA	NA
		60%	0.139	0.096	0.098	NA	NA
	0.5	20%	0.073	0.123	0.126	NA	NA
		40%	0.070	0.108	0.108	NA	NA
		60%	0.213	0.144	0.146	NA	NA
	0.8	20%	0.104	0.108	0.108	NA	NA
		40%	0.051	0.100	0.100	NA	NA
		60%	0.067	0.154	0.153	NA	NA
1	0.2	20%	0.110	0.118	0.119	NA	NA
		40%	0.041	0.075	0.080	NA	NA
		60%	0.044	0.047	0.050	NA	NA
	0.5	20%	0.099	0.124	0.124	NA	NA
		40%	0.047	0.171	0.175	NA	NA
		60%	0.042	0.119	0.124	NA	NA
	0.8	20%	0.093	0.080	0.079	NA	NA
		40%	0.071	0.129	0.132	NA	NA
		60%	0.231	0.228	0.231	NA	NA

Table 30
Bias for Intercept under Each Study Condition

Manipulated Factors			Bias				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.083	-0.018	-0.047	NA	NA
		40%	0.066	0.027	-0.001	NA	NA
		60%	-0.001	0.035	0.008	NA	NA
	0.5	20%	0.069	0.389	0.357	NA	NA
		40%	0.123	0.066	0.038	NA	NA
		60%	0.214	0.273	0.230	NA	NA
	0.8	20%	0.086	0.032	0.008	NA	NA
		40%	0.145	0.003	-0.023	NA	NA
		60%	0.116	0.084	0.053	NA	NA
1	0.2	20%	0.021	-0.075	-0.106	NA	NA
		40%	0.000	-0.070	-0.095	NA	NA
		60%	-0.027	0.053	0.013	NA	NA
	0.5	20%	0.041	-0.079	-0.103	NA	NA
		40%	0.027	-0.002	-0.044	NA	NA
		60%	-0.004	-0.121	-0.145	NA	NA
	0.8	20%	0.058	-0.093	-0.114	NA	NA
		40%	0.057	-0.165	-0.181	NA	NA
		60%	-0.290	-0.287	-0.299	NA	NA

Table 31
SE for Intercept under Each Study Condition

Manipulated Factors			SE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.096	0.084	0.080	NA	NA
		40%	0.039	0.034	0.031	NA	NA
		60%	0.041	0.050	0.048	NA	NA
	0.5	20%	0.050	0.045	0.051	NA	NA
		40%	0.052	0.054	0.050	NA	NA
		60%	0.052	0.046	0.042	NA	NA
	0.8	20%	0.098	0.082	0.077	NA	NA
		40%	0.042	0.029	0.028	NA	NA
		60%	0.042	0.036	0.034	NA	NA
1	0.2	20%	0.076	0.056	0.054	NA	NA
		40%	0.029	0.023	0.023	NA	NA
		60%	0.029	0.029	0.027	NA	NA
	0.5	20%	0.062	0.048	0.045	NA	NA
		40%	0.052	0.055	0.050	NA	NA
		60%	0.024	0.025	0.023	NA	NA
	0.8	20%	0.083	0.063	0.059	NA	NA
		40%	0.066	0.053	0.051	NA	NA
		60%	0.070	0.034	0.031	NA	NA

Table 32
RMSE for Intercept under Each Study Condition

Manipulated Factors			RMSE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.127	0.086	0.092	NA	NA
		40%	0.076	0.043	0.031	NA	NA
		60%	0.041	0.061	0.049	NA	NA
	0.5	20%	0.085	0.392	0.360	NA	NA
		40%	0.133	0.085	0.063	NA	NA
		60%	0.220	0.277	0.234	NA	NA
	0.8	20%	0.130	0.088	0.077	NA	NA
		40%	0.151	0.029	0.036	NA	NA
		60%	0.123	0.091	0.063	NA	NA
1	0.2	20%	0.079	0.094	0.119	NA	NA
		40%	0.029	0.074	0.098	NA	NA
		60%	0.040	0.061	0.030	NA	NA
	0.5	20%	0.074	0.092	0.112	NA	NA
		40%	0.058	0.055	0.067	NA	NA
		60%	0.025	0.123	0.147	NA	NA
	0.8	20%	0.101	0.112	0.129	NA	NA
		40%	0.087	0.174	0.188	NA	NA
		60%	0.298	0.289	0.300	NA	NA

Table 33
Bias for Testlet Effect Variance under Each Study Condition

Manipulated Factors			Bias				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.649	-0.048	NA	0.489	NA
		40%	0.347	-0.048	NA	0.651	NA
		60%	0.316	-0.060	NA	0.716	NA
	0.5	20%	0.325	-0.079	NA	0.392	NA
		40%	0.516	-0.097	NA	0.564	NA
		60%	0.251	-0.081	NA	0.300	NA
	0.8	20%	0.575	-0.132	NA	0.532	NA
		40%	0.818	-0.137	NA	0.685	NA
		60%	0.510	-0.131	NA	0.864	NA
1	0.2	20%	0.199	-0.516	NA	0.263	NA
		40%	0.100	-0.526	NA	0.493	NA
		60%	0.007	-0.527	NA	0.855	NA
	0.5	20%	0.516	-0.725	NA	0.645	NA
		40%	0.251	-0.695	NA	0.776	NA
		60%	0.130	-0.696	NA	0.852	NA
	0.8	20%	0.978	-0.865	NA	0.832	NA
		40%	0.581	-0.857	NA	0.651	NA
		60%	1.009	-0.870	NA	1.089	NA

Table 34

SE for Testlet Effect Variance under Each Study Condition

Manipulated Factors			SE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.189	0.033	NA	0.151	NA
		40%	0.128	0.034	NA	0.161	NA
		60%	0.166	0.032	NA	0.217	NA
	0.5	20%	0.085	0.024	NA	0.075	NA
		40%	0.138	0.020	NA	0.116	NA
		60%	0.080	0.026	NA	0.066	NA
	0.8	20%	0.099	0.012	NA	0.089	NA
		40%	0.121	0.012	NA	0.094	NA
		60%	0.100	0.016	NA	0.105	NA
1	0.2	20%	0.247	0.061	NA	0.252	NA
		40%	0.187	0.060	NA	0.276	NA
		60%	0.205	0.057	NA	0.300	NA
	0.5	20%	0.289	0.035	NA	0.317	NA
		40%	0.227	0.051	NA	0.357	NA
		60%	0.265	0.052	NA	0.509	NA
	0.8	20%	0.300	0.017	NA	0.231	NA
		40%	0.273	0.018	NA	0.211	NA
		60%	0.809	0.018	NA	0.303	NA

Table 35

RMSE for Testlet Effect Variance under Each Study Condition

Manipulated Factors			RMSE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.676	0.060	NA	0.512	NA
		40%	0.370	0.060	NA	0.671	NA
		60%	0.359	0.069	NA	0.749	NA
	0.5	20%	0.336	0.083	NA	0.399	NA
		40%	0.534	0.099	NA	0.575	NA
		60%	0.263	0.085	NA	0.307	NA
	0.8	20%	0.583	0.132	NA	0.539	NA
		40%	0.827	0.137	NA	0.692	NA
		60%	0.520	0.131	NA	0.871	NA
1	0.2	20%	0.318	0.519	NA	0.364	NA
		40%	0.222	0.530	NA	0.567	NA
		60%	0.225	0.530	NA	0.907	NA
	0.5	20%	0.592	0.726	NA	0.718	NA
		40%	0.339	0.698	NA	0.855	NA
		60%	0.298	0.698	NA	0.993	NA
	0.8	20%	1.023	0.866	NA	0.863	NA
		40%	0.642	0.858	NA	0.685	NA
		60%	1.295	0.870	NA	1.130	NA

Table 36

Bias for Dual Testlet Correlation under Each Study Condition

Manipulated Factors			Bias				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.448	NA	NA	0.404	NA
		40%	0.341	NA	NA	0.428	NA
		60%	0.324	NA	NA	0.454	NA
	0.5	20%	0.172	NA	NA	0.188	NA
		40%	0.230	NA	NA	0.235	NA
		60%	0.143	NA	NA	0.172	NA
	0.8	20%	0.046	NA	NA	0.040	NA
		40%	0.068	NA	NA	0.056	NA
		60%	0.033	NA	NA	0.063	NA
1	0.2	20%	0.049	NA	NA	0.065	NA
		40%	0.024	NA	NA	0.122	NA
		60%	-0.017	NA	NA	0.172	NA
	0.5	20%	0.076	NA	NA	0.089	NA
		40%	0.029	NA	NA	0.087	NA
		60%	-0.025	NA	NA	0.069	NA
	0.8	20%	0.035	NA	NA	0.026	NA
		40%	0.017	NA	NA	0.018	NA
		60%	-0.001	NA	NA	0.040	NA

Table 37
SE for Dual Testlet Correlation under Each Study Condition

Manipulated Factors			SE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.056	NA	NA	0.065	NA
		40%	0.086	NA	NA	0.067	NA
		60%	0.078	NA	NA	0.060	NA
	0.5	20%	0.054	NA	NA	0.060	NA
		40%	0.046	NA	NA	0.044	NA
		60%	0.065	NA	NA	0.052	NA
	0.8	20%	0.030	NA	NA	0.032	NA
		40%	0.023	NA	NA	0.026	NA
		60%	0.041	NA	NA	0.035	NA
1	0.2	20%	0.046	NA	NA	0.047	NA
		40%	0.045	NA	NA	0.035	NA
		60%	0.061	NA	NA	0.037	NA
	0.5	20%	0.026	NA	NA	0.024	NA
		40%	0.044	NA	NA	0.031	NA
		60%	0.049	NA	NA	0.032	NA
	0.8	20%	0.019	NA	NA	0.021	NA
		40%	0.028	NA	NA	0.028	NA
		60%	0.040	NA	NA	0.025	NA

Table 38

RMSE for Dual Testlet Correlation under Each Study Condition

Manipulated Factors			RMSE				
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL
0.5	0.2	20%	0.451	NA	NA	0.409	NA
		40%	0.352	NA	NA	0.434	NA
		60%	0.333	NA	NA	0.458	NA
	0.5	20%	0.180	NA	NA	0.197	NA
		40%	0.234	NA	NA	0.239	NA
		60%	0.157	NA	NA	0.180	NA
	0.8	20%	0.054	NA	NA	0.051	NA
		40%	0.072	NA	NA	0.062	NA
		60%	0.053	NA	NA	0.072	NA
1	0.2	20%	0.067	NA	NA	0.081	NA
		40%	0.051	NA	NA	0.127	NA
		60%	0.063	NA	NA	0.175	NA
	0.5	20%	0.081	NA	NA	0.092	NA
		40%	0.053	NA	NA	0.092	NA
		60%	0.055	NA	NA	0.076	NA
	0.8	20%	0.040	NA	NA	0.033	NA
		40%	0.033	NA	NA	0.033	NA
		60%	0.040	NA	NA	0.047	NA

Table 39

Bias for Overall Ability under Each Study Condition

Manipulated Factors			Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	-0.001	-0.005	-0.005	-0.004	-0.007	-0.058
		40%	0.000	0.000	0.001	-0.004	-0.007	-0.066
		60%	-0.026	-0.015	-0.015	-0.004	-0.008	-0.070
	0.5	20%	-0.001	-0.034	-0.035	-0.001	-0.005	-0.043
		40%	0.001	-0.005	-0.005	-0.005	-0.008	-0.086
		60%	-0.035	-0.015	-0.017	-0.009	-0.011	-0.130
	0.8	20%	0.002	-0.007	-0.007	-0.001	-0.005	-0.064
		40%	0.002	-0.004	-0.004	-0.005	-0.008	-0.073
		60%	-0.007	0.003	0.003	-0.007	-0.009	-0.129
1	0.2	20%	-0.001	-0.006	-0.006	-0.001	-0.005	-0.040
		40%	0.001	-0.024	-0.021	-0.002	-0.006	-0.062
		60%	-0.001	-0.043	-0.046	-0.004	-0.006	-0.081
	0.5	20%	0.000	-0.004	-0.004	-0.002	-0.005	-0.067
		40%	0.001	0.009	0.010	-0.004	-0.007	-0.101
		60%	0.000	-0.045	-0.045	-0.003	-0.006	-0.074
	0.8	20%	0.001	-0.002	-0.002	-0.002	-0.005	-0.043
		40%	-0.004	-0.026	-0.026	-0.003	-0.006	-0.104
		60%	0.010	0.018	0.018	-0.005	-0.007	-0.042

Table 40

SE for Overall Ability under Each Study Condition

Manipulated Factors			SE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.270	0.324	0.327	0.277	0.326	0.539
		40%	0.279	0.326	0.328	0.260	0.329	0.521
		60%	0.279	0.318	0.320	0.258	0.318	0.517
	0.5	20%	0.287	0.342	0.343	0.280	0.343	0.547
		40%	0.261	0.327	0.328	0.257	0.327	0.549
		60%	0.291	0.343	0.343	0.285	0.340	0.545
	0.8	20%	0.254	0.332	0.332	0.257	0.332	0.563
		40%	0.224	0.314	0.311	0.230	0.311	0.562
		60%	0.248	0.320	0.318	0.226	0.318	0.554
1	0.2	20%	0.239	0.302	0.297	0.235	0.298	0.515
		40%	0.242	0.305	0.299	0.227	0.300	0.501
		60%	0.251	0.297	0.301	0.213	0.302	0.502
	0.5	20%	0.209	0.289	0.289	0.205	0.290	0.544
		40%	0.224	0.297	0.295	0.208	0.296	0.536
		60%	0.233	0.289	0.293	0.215	0.294	0.506
	0.8	20%	0.175	0.273	0.270	0.175	0.270	0.558
		40%	0.197	0.287	0.284	0.197	0.284	0.563
		60%	0.212	0.277	0.276	0.194	0.275	0.588

Table 41

RMSE for Overall Ability under Each Study Condition

Manipulated Factors			RMSE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.476	0.470	0.472	0.472	0.472	0.767
		40%	0.465	0.475	0.476	0.469	0.477	0.713
		60%	0.461	0.465	0.466	0.470	0.464	0.720
	0.5	20%	0.490	0.503	0.504	0.489	0.502	0.747
		40%	0.483	0.494	0.495	0.483	0.495	0.794
		60%	0.493	0.509	0.509	0.489	0.507	0.753
	0.8	20%	0.499	0.515	0.514	0.498	0.514	0.840
		40%	0.491	0.503	0.501	0.488	0.501	0.905
		60%	0.488	0.509	0.507	0.493	0.508	0.844
1	0.2	20%	0.583	0.623	0.623	0.582	0.623	0.948
		40%	0.580	0.622	0.622	0.578	0.623	0.927
		60%	0.593	0.626	0.628	0.588	0.630	0.914
	0.5	20%	0.602	0.652	0.653	0.602	0.653	1.119
		40%	0.602	0.649	0.650	0.603	0.651	1.053
		60%	0.605	0.648	0.651	0.607	0.651	0.993
	0.8	20%	0.609	0.678	0.673	0.607	0.674	1.310
		40%	0.619	0.676	0.673	0.621	0.673	1.225
		60%	0.630	0.676	0.673	0.625	0.674	1.286

Table 42

Bias for Subscore of Addition under Each Study Condition

Manipulated Factors			Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.014	-0.011	-0.018	-0.007	-0.037	0.023
		40%	0.003	0.010	0.008	-0.005	-0.013	-0.013
		60%	-0.015	-0.106	-0.113	-0.100	-0.108	-0.066
	0.5	20%	0.021	0.072	0.059	-0.097	-0.114	0.012
		40%	-0.002	0.017	0.019	0.102	0.097	-0.033
		60%	-0.044	-0.050	-0.055	0.118	0.103	-0.056
	0.8	20%	0.031	0.011	0.004	-0.144	-0.156	0.161
		40%	0.013	-0.016	-0.021	-0.031	-0.038	-0.028
		60%	0.001	-0.030	-0.038	0.014	-0.010	-0.108
1	0.2	20%	0.010	-0.022	-0.034	-0.079	-0.101	-0.008
		40%	0.021	0.020	0.011	0.127	0.096	-0.040
		60%	-0.014	-0.050	-0.067	-0.062	-0.065	-0.049
	0.5	20%	0.003	-0.008	-0.016	-0.076	-0.083	-0.006
		40%	-0.001	-0.007	-0.002	0.077	0.074	-0.020
		60%	0.004	-0.102	-0.123	-0.157	-0.169	-0.061
	0.8	20%	0.013	-0.002	-0.010	-0.054	-0.065	-0.038
		40%	0.035	-0.034	-0.052	-0.073	-0.108	-0.170
		60%	-0.051	-0.105	-0.117	-0.104	-0.116	-0.147

Table 43

SE for Subscore of Addition under Each Study Condition

Manipulated Factors			SE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.757	0.722	0.716	0.762	0.727	1.072
		40%	0.665	0.634	0.657	0.679	0.657	0.883
		60%	0.586	0.564	0.560	0.613	0.584	0.778
	0.5	20%	0.761	0.741	0.738	0.769	0.734	1.087
		40%	0.652	0.622	0.638	0.671	0.653	0.905
		60%	0.571	0.582	0.578	0.625	0.620	0.820
	0.8	20%	0.718	0.687	0.679	0.730	0.695	0.998
		40%	0.696	0.644	0.656	0.696	0.659	1.006
		60%	0.634	0.631	0.623	0.651	0.624	0.932
1	0.2	20%	0.752	0.705	0.699	0.770	0.714	1.065
		40%	0.633	0.572	0.625	0.658	0.636	0.922
		60%	0.580	0.570	0.568	0.614	0.578	0.797
	0.5	20%	0.688	0.631	0.638	0.692	0.639	0.979
		40%	0.638	0.582	0.614	0.669	0.617	0.910
		60%	0.557	0.521	0.522	0.606	0.556	0.805
	0.8	20%	0.667	0.570	0.566	0.681	0.583	0.943
		40%	0.673	0.604	0.620	0.683	0.631	1.072
		60%	0.691	0.599	0.593	0.696	0.618	1.043

Table 44

RMSE for Subscore of Addition under Each Study Condition

Manipulated Factors			RMSE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.827	0.784	0.781	0.823	0.794	1.227
		40%	0.737	0.708	0.731	0.757	0.732	1.013
		60%	0.669	0.651	0.650	0.708	0.673	0.907
	0.5	20%	0.829	0.814	0.811	0.842	0.813	1.227
		40%	0.740	0.711	0.728	0.761	0.749	1.075
		60%	0.666	0.681	0.679	0.723	0.738	0.978
	0.8	20%	0.806	0.777	0.771	0.825	0.802	1.189
		40%	0.799	0.746	0.757	0.793	0.761	1.239
		60%	0.734	0.737	0.731	0.756	0.731	1.139
1	0.2	20%	0.885	0.863	0.868	0.904	0.890	1.334
		40%	0.800	0.764	0.825	0.829	0.844	1.219
		60%	0.763	0.773	0.781	0.795	0.794	1.092
	0.5	20%	0.859	0.839	0.854	0.865	0.857	1.378
		40%	0.827	0.812	0.850	0.856	0.857	1.274
		60%	0.765	0.764	0.776	0.815	0.828	1.175
	0.8	20%	0.865	0.834	0.831	0.874	0.857	1.449
		40%	0.857	0.842	0.859	0.867	0.880	1.551
		60%	0.879	0.836	0.834	0.888	0.867	1.514

Table 45

Bias for Subscore of Subtraction under Each Study Condition

Manipulated Factors			Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	-0.004	-0.030	-0.039	0.043	0.028	-0.144
		40%	0.003	-0.075	-0.085	0.024	0.004	-0.047
		60%	-0.036	-0.148	-0.158	-0.083	-0.103	-0.138
	0.5	20%	0.009	-0.091	-0.103	-0.164	-0.183	-0.048
		40%	0.010	-0.024	-0.026	0.081	0.067	-0.026
		60%	-0.003	-0.002	-0.013	0.106	0.090	-0.097
	0.8	20%	-0.024	-0.029	-0.028	0.092	0.073	-0.045
		40%	-0.014	-0.086	-0.086	0.105	0.090	-0.128
		60%	0.006	-0.084	-0.094	-0.034	-0.061	-0.125
1	0.2	20%	0.000	-0.067	-0.075	-0.165	-0.165	-0.011
		40%	-0.009	-0.225	-0.235	-0.181	-0.190	-0.112
		60%	-0.008	-0.099	-0.128	-0.104	-0.129	-0.101
	0.5	20%	-0.006	-0.048	-0.058	-0.020	-0.037	-0.093
		40%	-0.002	-0.057	-0.069	-0.015	-0.027	-0.050
		60%	-0.009	-0.175	-0.192	-0.175	-0.184	-0.065
	0.8	20%	-0.009	-0.076	-0.081	-0.007	-0.024	-0.064
		40%	-0.012	-0.130	-0.130	0.031	0.014	-0.050
		60%	-0.021	-0.100	-0.093	0.160	0.161	-0.006

Table 46

SE for Subscore of Subtraction under Each Study Condition

Manipulated Factors			SE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.701	0.652	0.649	0.703	0.669	1.019
		40%	0.622	0.600	0.598	0.637	0.621	0.867
		60%	0.625	0.592	0.588	0.657	0.628	0.881
	0.5	20%	0.789	0.755	0.748	0.774	0.749	1.084
		40%	0.590	0.564	0.558	0.612	0.593	0.809
		60%	0.686	0.658	0.665	0.663	0.652	0.930
	0.8	20%	0.702	0.664	0.656	0.718	0.675	0.983
		40%	0.647	0.590	0.582	0.665	0.623	0.922
		60%	0.691	0.667	0.659	0.696	0.660	1.011
1	0.2	20%	0.667	0.636	0.630	0.643	0.619	0.933
		40%	0.627	0.580	0.572	0.636	0.595	0.853
		60%	0.588	0.572	0.574	0.612	0.592	0.910
	0.5	20%	0.702	0.633	0.623	0.704	0.632	1.032
		40%	0.579	0.553	0.547	0.596	0.555	0.820
		60%	0.573	0.516	0.517	0.624	0.559	0.811
	0.8	20%	0.649	0.570	0.563	0.659	0.587	1.012
		40%	0.611	0.547	0.541	0.641	0.575	0.897
		60%	0.573	0.493	0.490	0.612	0.538	0.891

Table 47

RMSE for Subscore of Subtraction under Each Study Condition

Manipulated Factors			RMSE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.778	0.721	0.721	0.773	0.742	1.198
		40%	0.699	0.678	0.679	0.719	0.699	1.003
		60%	0.703	0.681	0.680	0.743	0.710	1.004
	0.5	20%	0.854	0.826	0.821	0.857	0.835	1.216
		40%	0.689	0.661	0.656	0.708	0.696	0.986
		60%	0.764	0.741	0.749	0.754	0.741	1.074
	0.8	20%	0.792	0.757	0.750	0.808	0.774	1.176
		40%	0.759	0.701	0.695	0.769	0.740	1.159
		60%	0.780	0.765	0.759	0.797	0.757	1.202
1	0.2	20%	0.830	0.822	0.828	0.824	0.825	1.246
		40%	0.800	0.799	0.802	0.825	0.822	1.160
		60%	0.764	0.772	0.786	0.792	0.810	1.205
	0.5	20%	0.867	0.839	0.837	0.870	0.846	1.422
		40%	0.787	0.795	0.799	0.802	0.806	1.210
		60%	0.777	0.767	0.779	0.830	0.831	1.176
	0.8	20%	0.850	0.827	0.825	0.854	0.854	1.544
		40%	0.822	0.815	0.813	0.844	0.853	1.380
		60%	0.801	0.778	0.778	0.846	0.859	1.439

Table 48

Bias for Subscore of Multiplication under Each Study Condition

Manipulated Factors			Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	-0.021	-0.035	-0.035	0.181	0.158	-0.083
		40%	-0.006	-0.124	-0.128	0.039	0.015	-0.114
		60%	-0.100	-0.139	-0.150	-0.207	-0.213	-0.093
	0.5	20%	-0.023	0.007	0.002	0.077	0.060	-0.027
		40%	0.006	-0.168	-0.172	0.030	0.009	-0.041
		60%	0.108	0.075	0.060	0.112	0.096	-0.039
	0.8	20%	-0.001	-0.019	-0.025	0.006	-0.014	-0.045
		40%	-0.002	-0.104	-0.106	0.016	0.000	-0.114
		60%	0.047	-0.119	-0.123	0.038	0.017	-0.141
1	0.2	20%	-0.016	-0.087	-0.091	0.125	0.109	-0.097
		40%	-0.030	-0.162	-0.167	-0.071	-0.073	-0.082
		60%	-0.011	-0.158	-0.182	-0.072	-0.092	-0.072
	0.5	20%	-0.016	-0.050	-0.051	0.025	0.027	-0.024
		40%	0.004	-0.063	-0.079	-0.043	-0.090	-0.285
		60%	-0.041	-0.118	-0.121	0.015	0.025	-0.013
	0.8	20%	-0.030	-0.066	-0.065	0.101	0.089	-0.118
		40%	-0.014	-0.148	-0.140	0.112	0.093	-0.057
		60%	-0.240	-0.228	-0.236	-0.182	-0.183	-0.255

Table 49

SE for Subscore of Multiplication under Each Study Condition

Manipulated Factors			SE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.680	0.632	0.628	0.698	0.662	0.930
		40%	0.679	0.643	0.653	0.698	0.685	1.025
		60%	0.657	0.647	0.643	0.635	0.615	0.916
	0.5	20%	0.723	0.713	0.710	0.717	0.694	0.981
		40%	0.700	0.642	0.653	0.726	0.702	1.053
		60%	0.639	0.655	0.647	0.649	0.639	0.864
	0.8	20%	0.734	0.699	0.689	0.741	0.698	1.068
		40%	0.686	0.612	0.618	0.675	0.631	0.990
		60%	0.611	0.576	0.572	0.613	0.597	0.908
1	0.2	20%	0.710	0.661	0.652	0.743	0.697	0.993
		40%	0.600	0.591	0.602	0.601	0.595	0.859
		60%	0.593	0.558	0.555	0.629	0.595	0.864
	0.5	20%	0.630	0.581	0.576	0.652	0.598	0.899
		40%	0.696	0.658	0.661	0.707	0.668	1.195
		60%	0.511	0.493	0.491	0.562	0.520	0.732
	0.8	20%	0.672	0.579	0.571	0.691	0.606	0.997
		40%	0.637	0.570	0.578	0.631	0.590	0.959
		60%	0.557	0.500	0.497	0.578	0.514	0.851

Table 50
RMSE for Subscore of Multiplication under Each Study Condition

Manipulated Factors			RMSE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.762	0.704	0.702	0.787	0.753	1.088
		40%	0.749	0.722	0.735	0.773	0.759	1.176
		60%	0.735	0.732	0.732	0.749	0.719	1.079
	0.5	20%	0.796	0.790	0.788	0.796	0.773	1.117
		40%	0.780	0.737	0.749	0.800	0.785	1.212
		60%	0.729	0.754	0.747	0.738	0.745	1.023
	0.8	20%	0.820	0.786	0.777	0.823	0.786	1.244
		40%	0.791	0.723	0.728	0.779	0.736	1.228
		60%	0.714	0.691	0.688	0.724	0.708	1.132
1	0.2	20%	0.854	0.835	0.837	0.890	0.891	1.266
		40%	0.783	0.809	0.831	0.786	0.812	1.194
		60%	0.770	0.764	0.772	0.803	0.810	1.150
	0.5	20%	0.817	0.804	0.807	0.833	0.836	1.308
		40%	0.857	0.861	0.874	0.870	0.885	1.611
		60%	0.735	0.752	0.760	0.773	0.793	1.110
	0.8	20%	0.867	0.834	0.829	0.881	0.877	1.519
		40%	0.841	0.829	0.838	0.839	0.851	1.451
		60%	0.830	0.791	0.792	0.825	0.808	1.357

Table 51
Bias for Subscore of Division under Each Study Condition

Manipulated Factors			Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	-0.027	-0.073	-0.082	0.038	0.032	0.009
		40%	0.002	-0.100	-0.113	-0.084	-0.108	-0.043
		60%	-0.237	-0.287	-0.291	-0.220	-0.234	-0.213
	0.5	20%	-0.024	-0.114	-0.113	0.043	0.029	-0.031
		40%	0.025	-0.060	-0.083	-0.126	-0.163	-0.264
		60%	0.198	-0.067	-0.070	0.163	0.148	0.080
	0.8	20%	0.011	-0.125	-0.133	0.057	0.022	-0.190
		40%	0.023	-0.088	-0.095	-0.084	-0.102	-0.002
		60%	0.025	-0.015	-0.017	0.101	0.080	-0.050
1	0.2	20%	-0.022	-0.096	-0.100	0.002	-0.003	-0.001
		40%	-0.013	-0.153	-0.169	-0.151	-0.171	-0.153
		60%	-0.019	-0.079	-0.090	-0.008	-0.011	-0.034
	0.5	20%	0.001	-0.084	-0.099	0.053	0.019	-0.180
		40%	0.002	-0.104	-0.125	0.016	-0.034	-0.194
		60%	-0.012	-0.172	-0.184	-0.115	-0.128	-0.034
	0.8	20%	0.019	-0.071	-0.082	-0.229	-0.246	0.007
		40%	0.000	-0.123	-0.134	-0.039	-0.066	-0.163
		60%	-0.262	-0.208	-0.212	-0.079	-0.080	-0.267

Table 52
SE for Subscore of Division under Each Study Condition

Manipulated Factors			SE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.668	0.610	0.605	0.721	0.681	0.940
		40%	0.669	0.646	0.673	0.683	0.693	0.966
		60%	0.659	0.620	0.622	0.663	0.616	0.881
	0.5	20%	0.697	0.666	0.664	0.726	0.702	0.978
		40%	0.688	0.664	0.681	0.696	0.696	1.211
		60%	0.670	0.574	0.570	0.673	0.656	0.888
	0.8	20%	0.726	0.679	0.673	0.740	0.713	1.163
		40%	0.578	0.544	0.554	0.624	0.612	0.912
		60%	0.643	0.621	0.614	0.663	0.626	0.905
1	0.2	20%	0.635	0.602	0.593	0.638	0.601	0.905
		40%	0.581	0.581	0.604	0.585	0.609	0.942
		60%	0.566	0.541	0.545	0.598	0.560	0.761
	0.5	20%	0.669	0.617	0.615	0.688	0.643	1.078
		40%	0.625	0.605	0.636	0.639	0.649	1.062
		60%	0.612	0.557	0.553	0.658	0.596	0.867
	0.8	20%	0.703	0.591	0.585	0.717	0.609	1.012
		40%	0.638	0.578	0.599	0.632	0.597	1.003
		60%	0.562	0.513	0.510	0.586	0.527	0.867

Table 53
RMSE for Subscore of Division under Each Study Condition

Manipulated Factors			RMSE					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.753	0.684	0.681	0.786	0.759	1.079
		40%	0.748	0.726	0.760	0.771	0.781	1.119
		60%	0.782	0.748	0.750	0.796	0.723	1.018
	0.5	20%	0.773	0.752	0.750	0.798	0.785	1.116
		40%	0.774	0.748	0.771	0.788	0.799	1.458
		60%	0.773	0.667	0.664	0.769	0.768	1.054
	0.8	20%	0.811	0.771	0.768	0.821	0.801	1.376
		40%	0.706	0.670	0.679	0.733	0.749	1.183
		60%	0.742	0.726	0.719	0.771	0.738	1.109
1	0.2	20%	0.802	0.797	0.799	0.805	0.805	1.205
		40%	0.767	0.799	0.855	0.780	0.862	1.335
		60%	0.755	0.752	0.764	0.783	0.780	1.057
	0.5	20%	0.836	0.827	0.834	0.853	0.866	1.498
		40%	0.810	0.828	0.875	0.825	0.887	1.494
		60%	0.800	0.787	0.792	0.845	0.839	1.202
	0.8	20%	0.891	0.838	0.835	0.922	0.892	1.506
		40%	0.842	0.832	0.854	0.843	0.846	1.489
		60%	0.834	0.803	0.803	0.816	0.815	1.421

Appendix D SD of Bias for Overall Ability and Subscores

Table 54

SD of Bias for Overall Ability

Manipulated Factors			SD of Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.450	0.380	0.379	0.435	0.379	0.600
		40%	0.425	0.384	0.381	0.452	0.381	0.538
		60%	0.414	0.375	0.375	0.450	0.374	0.563
	0.5	20%	0.448	0.406	0.405	0.453	0.404	0.561
		40%	0.467	0.415	0.414	0.472	0.414	0.635
		60%	0.453	0.420	0.419	0.457	0.419	0.564
	0.8	20%	0.499	0.445	0.443	0.496	0.443	0.688
		40%	0.508	0.439	0.438	0.498	0.438	0.793
		60%	0.489	0.444	0.443	0.518	0.443	0.706
1	0.2	20%	0.627	0.630	0.634	0.628	0.634	0.919
		40%	0.623	0.624	0.630	0.632	0.631	0.900
		60%	0.628	0.636	0.637	0.648	0.639	0.876
	0.5	20%	0.674	0.684	0.685	0.677	0.685	1.133
		40%	0.663	0.673	0.675	0.676	0.676	1.045
		60%	0.666	0.680	0.682	0.680	0.684	0.995
	0.8	20%	0.713	0.732	0.728	0.709	0.729	1.364
		40%	0.704	0.722	0.721	0.705	0.721	1.257
		60%	0.706	0.722	0.719	0.713	0.720	1.326

Table 55
SD of Bias for Subscore of Addition

Manipulated Factors			SD of Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.341	0.313	0.321	0.321	0.325	0.615
		40%	0.328	0.323	0.331	0.347	0.331	0.521
		60%	0.337	0.325	0.327	0.359	0.333	0.499
	0.5	20%	0.334	0.335	0.339	0.336	0.339	0.583
		40%	0.364	0.358	0.363	0.359	0.371	0.613
		60%	0.356	0.366	0.369	0.359	0.407	0.570
	0.8	20%	0.379	0.380	0.381	0.373	0.388	0.657
		40%	0.407	0.392	0.393	0.394	0.395	0.757
		60%	0.385	0.397	0.397	0.401	0.397	0.691
1	0.2	20%	0.491	0.525	0.545	0.491	0.555	0.853
		40%	0.525	0.546	0.579	0.526	0.588	0.863
		60%	0.535	0.565	0.581	0.540	0.591	0.816
	0.5	20%	0.549	0.597	0.612	0.551	0.613	1.045
		40%	0.563	0.613	0.635	0.568	0.639	0.965
		60%	0.570	0.606	0.621	0.571	0.656	0.946
	0.8	20%	0.597	0.667	0.669	0.589	0.687	1.193
		40%	0.568	0.636	0.643	0.567	0.657	1.194
		60%	0.578	0.624	0.626	0.581	0.651	1.169

Table 56
SD of Bias for Subscore of Subtraction

Manipulated Factors			SD of Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.347	0.314	0.319	0.328	0.327	0.651
		40%	0.327	0.318	0.321	0.343	0.329	0.532
		60%	0.331	0.318	0.318	0.352	0.327	0.506
	0.5	20%	0.334	0.332	0.334	0.343	0.336	0.571
		40%	0.372	0.360	0.362	0.364	0.377	0.596
		60%	0.347	0.351	0.354	0.357	0.351	0.560
	0.8	20%	0.381	0.378	0.379	0.372	0.389	0.677
		40%	0.411	0.388	0.389	0.389	0.409	0.734
		60%	0.376	0.380	0.380	0.402	0.380	0.673
1	0.2	20%	0.528	0.554	0.573	0.532	0.564	0.894
		40%	0.533	0.555	0.569	0.539	0.589	0.851
		60%	0.521	0.549	0.566	0.528	0.583	0.852
	0.5	20%	0.543	0.590	0.599	0.546	0.606	1.047
		40%	0.577	0.622	0.633	0.581	0.639	0.975
		60%	0.570	0.601	0.614	0.569	0.653	0.943
	0.8	20%	0.596	0.648	0.651	0.589	0.675	1.250
		40%	0.598	0.657	0.661	0.596	0.697	1.147
		60%	0.611	0.663	0.666	0.618	0.723	1.244

Table 57
SD of Bias for Subscore of Multiplication

Manipulated Factors			SD of Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.353	0.316	0.320	0.331	0.332	0.588
		40%	0.326	0.316	0.324	0.340	0.335	0.601
		60%	0.326	0.329	0.333	0.363	0.328	0.606
	0.5	20%	0.343	0.349	0.352	0.350	0.347	0.554
		40%	0.356	0.339	0.343	0.348	0.362	0.622
		60%	0.348	0.379	0.382	0.350	0.388	0.581
	0.8	20%	0.377	0.372	0.372	0.371	0.377	0.657
		40%	0.409	0.388	0.388	0.402	0.394	0.754
		60%	0.386	0.384	0.383	0.405	0.398	0.712
1	0.2	20%	0.503	0.535	0.553	0.504	0.583	0.838
		40%	0.543	0.579	0.604	0.546	0.597	0.910
		60%	0.525	0.543	0.555	0.529	0.587	0.823
	0.5	20%	0.561	0.605	0.615	0.559	0.637	1.037
		40%	0.534	0.586	0.601	0.540	0.609	1.123
		60%	0.581	0.620	0.635	0.582	0.668	0.932
	0.8	20%	0.593	0.651	0.653	0.584	0.688	1.222
		40%	0.593	0.646	0.652	0.593	0.666	1.177
		60%	0.634	0.642	0.644	0.621	0.667	1.139

Table 58
SD of Bias for Subscore of Division

Manipulated Factors			SD of Bias					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.358	0.311	0.312	0.318	0.345	0.548
		40%	0.344	0.330	0.350	0.364	0.361	0.596
		60%	0.371	0.331	0.330	0.412	0.322	0.498
	0.5	20%	0.345	0.343	0.345	0.340	0.361	0.566
		40%	0.364	0.349	0.361	0.362	0.370	0.825
		60%	0.349	0.349	0.350	0.350	0.390	0.591
	0.8	20%	0.373	0.358	0.360	0.363	0.377	0.746
		40%	0.427	0.400	0.402	0.395	0.445	0.806
		60%	0.385	0.392	0.392	0.399	0.400	0.680
1	0.2	20%	0.527	0.555	0.571	0.527	0.578	0.866
		40%	0.536	0.574	0.636	0.540	0.641	1.019
		60%	0.540	0.564	0.578	0.545	0.591	0.804
	0.5	20%	0.536	0.583	0.594	0.535	0.621	1.104
		40%	0.556	0.599	0.636	0.564	0.649	1.118
		60%	0.559	0.586	0.597	0.561	0.634	0.911
	0.8	20%	0.589	0.643	0.645	0.580	0.670	1.198
		40%	0.594	0.642	0.650	0.603	0.649	1.176
		60%	0.626	0.652	0.655	0.618	0.685	1.209

Appendix E Identified Significant Effects for Subscore of Addition, Subscore of Subtraction and Subscore of Division

Table 59

ANOVA Results of Significant Effects on the Bias of $\hat{\theta}_{jA}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model * percent_dbcd	117.218	<0.001	0.013

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 60

ANOVA Results of Significant Effects on the SE of $\hat{\theta}_{jA}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	134587.285	<0.001	0.882
model * testlet.sd	1119.058	<0.001	0.059
model * dbcorr	1101.647	<0.001	0.109
model * percent_dbcd	650.056	<0.001	0.067
model * testlet.sd * dbcorr	221.608	<0.001	0.024
model * testlet.sd * percent_dbcd	145.601	<0.001	0.016
model * dbcorr * percent_dbcd	424.443	<0.001	0.086
model * testlet.sd * dbcorr * percent_dbcd	49.376	<0.001	0.011
Between			
testlet.sd	232.658	<0.001	0.013
percent_dbcd	1251.945	<0.001	0.122
testlet.sd * percent_dbcd	93.437	<0.001	0.01
dbcorr * percent_dbcd	252.546	<0.001	0.053

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 61

ANOVA Results of Significant Effects on the RMSE of $\hat{\theta}_{jA}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	29716.447	<0.001	0.623
model * testlet.sd	769.269	<0.001	0.041
model * dbcorr	500.254	<0.001	0.053
model * percent_dbcd	127.663	<0.001	0.014
model * testlet.sd *dbcorr	88.066	<0.001	0.01
model * dbcorr * percent_dbcd	70.484	<0.001	0.015
Between			
testlet.sd	1254.99	<0.001	0.065
dbcorr	154.421	<0.001	0.017
percent_dbcd	311.614	<0.001	0.033
dbcorr * percent_dbcd	54.47	<0.001	0.012

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 62

ANOVA Results of Significant Effects on the Bias of $\hat{\theta}_{jS}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	300.771	<0.001	0.016
model * testlet.sd * dbcorr * percent_dbcd	59.742	<0.001	0.013

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 63.

ANOVA Results of Significant Effects on SE of $\hat{\theta}_{JS}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	110395.684	<0.001	0.86
Model * testlet.sd	370.559	<0.001	0.02
model * dbcorr	462.529	<0.001	0.049
model * percent_dbcd	610.398	<0.001	0.064
model * testlet.sd * dbcorr	160.081	<0.001	0.017
model * dbcorr * percent_dbcd	79.272	<0.001	0.017
model * testlet.sd *dbcorr * percent_dbcd	166.046	<0.001	0.036
Between			
testlet.sd	1020.012	<0.001	0.054
percent_dbcd	869.946	<0.001	0.088
testlet.sd * percent_dbcd	115.409	<0.001	0.013
dbcorr * percent_dbcd	118.45	<0.001	0.026

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 64

ANOVA results of Significant Effects on the RMSE of $\hat{\theta}_{JS}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	27037.106	<0.001	0.601
model * testlet.sd	587.772	<0.001	0.032
model * dbcorr	315.537	<0.001	0.034
model * percent_dbcd	130.408	<0.001	0.014
model * testlet.sd * dbcorr	87.94	<0.001	0.01
Between			
testlet.sd	865.562	<0.001	0.046
dbcorr	111.203	<0.001	0.012
percent_dbcd	187.975	<0.001	0.02

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 65

ANOVA results of Significant Effects on the Bias of $\hat{\theta}_{jD}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	359.975	<0.001	0.020

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 66

ANOVA results of Significant Effects on the SE of $\hat{\theta}_{jD}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	97562.138	<0.001	0.844
model * dbcorr	794.199	<0.001	0.081
model * percent_dbcd	1455.947	<0.001	0.139
model * testlet.sd * dbcorr	91.4	<0.001	0.01
model * dbcorr * percent_dbcd	293.368	<0.001	0.061
model * testlet.sd * dbcorr * percent_dbcd	332.139	<0.001	0.069
Between			
testlet.sd	657.552	<0.001	0.035
dbcorr	148.818	<0.001	0.016
percent_dbcd	530.319	<0.001	0.056
dbcorr * percent_dbcd	93.871	<0.001	0.02
testlet.sd * dbcorr * percent_dbcd	76.557	<0.001	0.017

Note: The testlet effect SD is shorted as "testlet.sd". The dual testlets correlation is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Table 67

ANOVA Results of Significant Effects on the RMSE of $\hat{\theta}_{jD}$

Source	F Value	p-value	η_p^2
Within (Huyhn-Feldt Adjustment)			
model	28662.802	<0.001	0.614
model * testlet.sd	494.56	<0.001	0.027
model * dbcorr	417.428	<0.001	0.044
model * percent_dbcd	412.191	<0.001	0.044
model * dbcorr * percent_dbcd	73.434	<0.001	0.016
model * testlet.sd * dbcorr * percent_dbcd	89.14	<0.001	0.019
Between			
testlet.sd	813.256	<0.001	0.043
dbcorr	117.366	<0.001	0.013
percent_dbcd	116.346	<0.001	0.013

Note: The correlation between testlet effects for dual testlets is abbreviated as "dbcorr". The percentage of double-coded items is abbreviated as "percent_dbcd".

Appendix F Reliability for Overall Ability and Subscores

Table 68

Reliability for Overall Ability

Manipulated Factors			Reliability					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.561	0.817	0.832	0.623	0.832	0.885
		40%	0.682	0.818	0.833	0.563	0.833	0.890
		60%	0.717	0.832	0.846	0.588	0.850	0.900
	0.5	20%	0.627	0.796	0.812	0.595	0.812	0.876
		40%	0.606	0.826	0.839	0.585	0.840	0.893
		60%	0.655	0.788	0.807	0.634	0.813	0.881
	0.8	20%	0.524	0.818	0.831	0.545	0.830	0.884
		40%	0.515	0.854	0.864	0.572	0.864	0.905
		60%	0.624	0.840	0.853	0.458	0.854	0.903
1	0.2	20%	0.667	0.849	0.865	0.647	0.865	0.912
		40%	0.691	0.852	0.867	0.573	0.867	0.913
		60%	0.720	0.842	0.871	0.446	0.870	0.913
	0.5	20%	0.549	0.871	0.886	0.503	0.886	0.922
		40%	0.621	0.864	0.879	0.436	0.880	0.919
		60%	0.657	0.861	0.876	0.414	0.876	0.916
	0.8	20%	0.389	0.898	0.908	0.443	0.908	0.931
		40%	0.480	0.889	0.896	0.450	0.896	0.926
		60%	0.436	0.890	0.904	0.365	0.906	0.930

Table 69
Reliability for Subscore of Addition

Manipulated Factors			Reliability					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.105	0.523	0.567	0.211	0.572	0.644
		40%	0.520	0.672	0.701	0.399	0.701	0.749
		60%	0.636	0.748	0.770	0.519	0.773	0.807
	0.5	20%	0.198	0.448	0.500	0.172	0.527	0.654
		40%	0.471	0.703	0.725	0.468	0.728	0.759
		60%	0.558	0.687	0.713	0.570	0.740	0.789
	0.8	20%	0.311	0.637	0.660	0.356	0.670	0.702
		40%	0.325	0.700	0.725	0.394	0.727	0.755
		60%	0.553	0.752	0.773	0.425	0.775	0.812
1	0.2	20%	0.188	0.510	0.585	0.187	0.594	0.680
		40%	0.538	0.714	0.748	0.429	0.744	0.805
		60%	0.617	0.743	0.795	0.394	0.797	0.824
	0.5	20%	0.375	0.732	0.759	0.333	0.760	0.785
		40%	0.453	0.735	0.759	0.300	0.762	0.792
		60%	0.539	0.769	0.796	0.386	0.808	0.830
	0.8	20%	0.077	0.738	0.762	0.186	0.769	0.767
		40%	0.304	0.768	0.786	0.294	0.791	0.810
		60%	0.184	0.747	0.791	0.153	0.794	0.796

Table 70
Reliability for Subscore of Subtraction

Manipulated Factors			Reliability					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.299	0.626	0.657	0.382	0.660	0.735
		40%	0.584	0.725	0.750	0.479	0.750	0.796
		60%	0.614	0.731	0.757	0.501	0.762	0.803
	0.5	20%	0.357	0.562	0.594	0.314	0.592	0.681
		40%	0.484	0.721	0.740	0.483	0.746	0.777
		60%	0.511	0.652	0.688	0.465	0.684	0.757
	0.8	20%	0.284	0.628	0.654	0.319	0.657	0.706
		40%	0.314	0.705	0.731	0.394	0.734	0.761
		60%	0.464	0.705	0.733	0.309	0.732	0.779
1	0.2	20%	0.547	0.729	0.764	0.506	0.763	0.805
		40%	0.532	0.718	0.756	0.420	0.757	0.796
		60%	0.639	0.760	0.805	0.432	0.812	0.850
	0.5	20%	0.285	0.690	0.731	0.230	0.732	0.769
		40%	0.566	0.790	0.813	0.427	0.814	0.836
		60%	0.470	0.732	0.766	0.277	0.777	0.808
	0.8	20%	0.299	0.788	0.813	0.375	0.819	0.819
		40%	0.341	0.785	0.801	0.367	0.806	0.814
		60%	0.334	0.803	0.833	0.394	0.838	0.835

Table 71
Reliability for Subscore of Multiplication

Manipulated Factors			Reliability					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.321	0.636	0.665	0.403	0.666	0.718
		40%	0.500	0.673	0.709	0.370	0.704	0.776
		60%	0.631	0.742	0.764	0.540	0.781	0.813
	0.5	20%	0.444	0.627	0.653	0.405	0.654	0.716
		40%	0.287	0.617	0.658	0.284	0.661	0.725
		60%	0.564	0.669	0.698	0.542	0.723	0.775
	0.8	20%	0.238	0.601	0.628	0.267	0.631	0.693
		40%	0.306	0.683	0.715	0.355	0.713	0.751
		60%	0.614	0.789	0.807	0.493	0.805	0.837
1	0.2	20%	0.302	0.586	0.648	0.306	0.655	0.710
		40%	0.588	0.749	0.783	0.475	0.780	0.822
		60%	0.603	0.743	0.792	0.381	0.791	0.828
	0.5	20%	0.382	0.742	0.777	0.369	0.782	0.796
		40%	0.404	0.702	0.752	0.212	0.752	0.816
		60%	0.572	0.785	0.810	0.410	0.822	0.838
	0.8	20%	0.200	0.757	0.789	0.304	0.796	0.790
		40%	0.396	0.785	0.805	0.339	0.811	0.819
		60%	0.346	0.815	0.840	0.350	0.846	0.847

Table 72
Reliability for Subscore of Division

Manipulated Factors			Reliability					
Testlet Effect SD	Dual Testlets Correlation	Percentage of Double-Coded Items	DT-MIRID	T-MIRID	MIRID	DTM	2PL	NCS
0.5	0.2	20%	0.192	0.610	0.616	0.335	0.631	0.689
		40%	0.503	0.653	0.690	0.401	0.695	0.755
		60%	0.515	0.650	0.682	0.386	0.707	0.772
	0.5	20%	0.440	0.680	0.652	0.425	0.661	0.723
		40%	0.382	0.646	0.681	0.397	0.698	0.799
		60%	0.485	0.658	0.687	0.481	0.682	0.731
	0.8	20%	0.307	0.615	0.679	0.332	0.676	0.755
		40%	0.341	0.726	0.752	0.460	0.767	0.794
		60%	0.497	0.725	0.748	0.364	0.750	0.784
1	0.2	20%	0.521	0.727	0.737	0.505	0.741	0.796
		40%	0.586	0.766	0.799	0.477	0.801	0.843
		60%	0.614	0.738	0.787	0.379	0.790	0.820
	0.5	20%	0.355	0.695	0.773	0.330	0.772	0.808
		40%	0.411	0.718	0.772	0.228	0.770	0.816
		60%	0.475	0.731	0.765	0.282	0.773	0.797
	0.8	20%	0.043	0.706	0.742	0.154	0.754	0.751
		40%	0.383	0.782	0.801	0.375	0.800	0.821
		60%	0.412	0.829	0.854	0.436	0.860	0.859

Appendix G Item Structure for Subscores of Addition, Multiplication and Division

Table 73

Information for Items in Subscore of Addition

	20% of Double-coded Items			40% of Double-coded Items			60% of Double-coded Items		
Position in subscore	Position in Test	Composite Item	Item Difficulty Used	Position in Test	Composite Item	Item difficulty Used	Position in Test	Composite Item	Item Difficult Used
1	1	no	1	1	no	1	1	no	1
2	2	no	2	2	no	2	5	yes	1
3	9	yes	1	7	yes	1	6	yes	1
4	10	yes	2	8	yes	2	7	yes	1
5	11	no	11	9	yes	1	11	no	11
6	12	no	12	10	yes	2	15	yes	11
7	19	yes	19	11	no	11	16	yes	11
8	21	no	21	12	no	12	17	yes	11
9	22	no	22	17	yes	11	21	no	21
10				18	yes	12	24	no	21
11							25	yes	21
12							26	yes	21

Note: 1. For component items, numbers in "Item Difficulty Used" are positions of the items in the test. For a composite item, the number in column 4, 7 and 10 is the position of the component item that assesses addition.

2. The scale of shades indicates which testlet the item belongs to, under the true condition. The lightest shade indicates the first testlet, the medium shade indicates the second testlet and the darkest indicates items for the dual testlets.

Table 74

Information for Items in Subscore of Multiplication

	20% of Double-coded Items			40% of Double-coded Items			60% of Double-coded Items		
Position in subscore	Position in Test	Composite Item	Item Difficulty Used	Position in Test	Composite Item	Item difficulty Used	Position in Test	Composite Item	Item Difficult Used
1	5	no	5	5	no	3	2	no	3
2	6	no	6	6	no	4	6	yes	3
3	10	yes	5	9	yes	3	8	yes	3
4	15	no	15	10	yes	4	10	yes	3
5	16	no	16	23	no	13	13	no	13
6	20	yes	15	24	no	14	16	yes	13
7	25	no	25	27	yes	13	18	yes	13
8	26	no	26	28	yes	14	20	yes	13
9	30	yes	25	29	yes	21	23	no	23
10				30	yes	22	26	yes	23
11							28	yes	23
12							30	yes	23

Note: 1. For component items, numbers in "Item Difficulty Used" are positions of the items in the test. For a composite item, the number in column 4, 7 and 10 is the position of the component item that assesses multiplication.

2. The scale of shades indicates which testlet the item belongs to, under the true condition. The lightest shade indicates the first testlet, the medium shade indicates the second testlet and the darkest indicates items for the dual testlets.

Table 75

Information for Items in Subscore of Division

	20% of Double-coded Items			40% of Double-coded Items			60% of Double-coded Items		
Position in subscore	Position in Test	Composite Item	Item Difficulty Used	Position in Test	Composite Item	Item difficulty Used	Position in Test	Composite Item	Item Difficult Used
1	7	no	7	15	no	15	4	no	4
2	8	no	8	16	no	16	7	yes	4
3	17	no	17	17	yes	15	9	yes	4
4	18	no	18	18	yes	16	10	yes	4
5	19	yes	17	19	yes	15	14	no	14
6	27	no	27	20	yes	16	17	yes	14
7	28	no	28	25	no	25	19	yes	14
8	29	yes	27	26	no	26	20	yes	14
9	30	yes	28	29	yes	25	24	no	24
10				30	yes	26	27	yes	24
11							29	yes	24
12							30	yes	24

Note: 1. For component items, numbers in column "Item Difficulty Used" are positions of the items in the test. For a composite item, the number in column 4, 7 and 10 is the position of the component item that assesses division.

2. The scale of shades indicates which testlet the item belongs to, under the true condition. The lightest shade indicates the first testlet, the medium shade indicates the second testlet and the darkest indicates items for the dual testlets.

References

Ackerman, T. A. (1987). The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence. ACT research report series, 87-14. Iowa City, Iowa: American College Testing Program.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278.

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Englhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, *21*(1), 1-23.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings, 2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Andersen. E. B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms?. *Medical Care*, 42(1), I-7.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153-169.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bechger, T., Verstralen, H., & Verhelst, N. (2002). Equivalent linear logistic test models. *Psychometrika, 67*(1), 123-136.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In L. A. Shepard & K. Ryan (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Boughton, K. A., Yao, L., & Lewis, D. M. (2006). *Reporting diagnostic subscale scores for tests composed of complex structure*. In Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153-168.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434-455.

Bunday, B. D. (1984). *Basic optimization methods*. London: Arnold.

Butter, R. P. (1994). *Item response models with internal restrictions on item difficulty.* Unpublished doctoral dissertation, Catholic University of Leuven, Belgium.

Butter, R., De Boeck, P., & Verhelst, N. (1998). An item response model with internal restrictions on item difficulty. *Psychometrika*, *63*(1), 47–63.

Camilli, G. (1994). Origin of the scaling constant d=1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics*, *19*(3), 293-295.

Carlson, J. E. (1993). *Dimensionality of NAEP instruments that incorporate polytomously scored items*. Paper presented at the annual meeting of the American Educational Research Association. Atlanta, GA.

Carlson, J. E., & Jirele, T. (1992, April). *Dimensionality of 1990 NAEP mathematics data*. Paper presented at the meeting of the American Educational Research Association. San Francisco, CA.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*(2), 1-37.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B.Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

College Board (2015). Test specifications for the redesigned SAT® (Technical Report). Retrieved from https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

De Boeck. P. (1991). *Componential IRT models.* Unpublished manuscript, University of Leuven, Belgium.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295–311.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620-639.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, *35*(4), 296-316.

DeMars, C. E. (2005). *Scoring subscales using multidimensional Item Response Theory models*. Paper presented at the Annual Meeting of the American Psychological Association. Washington, DC.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society.* 39 (Series B), 1-38.

Du, Z. (1998). *Modeling conditional item dependencies with a three-parameter logistic testlet model*. Unpublished doctoral dissertation. Columbia University, New York, NY.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006, April). *A comparison of subscale score augmentation methods using empirical data*. In annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics*, *31*(3), 241-259.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika*, *78*(1), 14-36.

Feinberg, R. A., & Wainer, H. (2014). When can we improve subscores by making them shorter?: The case against subscores with overlapping items. *Educational Measurement: Issues and Practice*, *33*(3), 47–54.

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education*, *10*(2), 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement*, *36*(2), 119-140.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Aeta Psychologica,37*(6)*,* 359-374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints*. Psychometrika, 48*(1)*, 3-26.*

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian data analysis. London: Chapman & Hall.

Gessaroli, M. E. (2004). *Using hierarchical multidimensional item response theory to estimate augmented subscores*. In annual meeting of the National Council on Measurement in Education. San Diego, CA.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Gill, P . E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. London: Academic Press.

Glass G.V., Peckham, P. D., & Sanders J. (1972) Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariances. *Review of Educational Research*, *42*(3), 237-288.

Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*, *216*(2), 89-101.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2), 57-63.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*(2), 101-125.

Hinkle, D., Wiersma, W., & Jurs, S. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston, MA: Houghton Mifflin.

Ip, E. H. S. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*(1), 73-91.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82-100.

Jiao, H., & Lissitz, R. (2014, October). *Exploring psychometric models for calibrating innovative items embedded in situations*. Paper presented at the Fourteenth Annual Maryland Assessment Conference: *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective*. University of Maryland, College Park.

Jiao, H., Lissitz, R. W., & Zhan, P. (2017). A non-compensatory testlet model for calibrating items embedded in multiple contexts. In H. Jiao & R.W. Lissitz (Eds.), *Technology-enhanced innovative assessment: Development, modeling, scoring from an interdisciplinary perspective*. Charlotte, NC: Information Age Publishing.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*(3), 311-321.

Kaplan, D. (1995). The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics*, *20*(1), 69-82.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218.

Lau, A.C., Jiao, H., & Lam, W. (April, 2004). *A simulation study to investigate the properties of pattern scoring and number-correct scoring using IRT model*. Paper presented in the annual meeting of the American Educational Research Association, San Diego, CA.

Lau, A.C., Jiao, H.,& Lam, W. (April, 2006). *A simulation study to compare pattern scoring and number-correct scoring with 3PL-IRT model*. Paper presented in the National Council on Measurement in Education Annual Meeting, San Francisco, CA.

Law, K. S., Wong, C. S., & Mobley, W. M. (1998). Toward a taxonomy of multidimensional constructs. *Academy of Management Review*, *23*(4), 741-755.

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. *Mixed-format tests: Psychometric Properties with a Primary Focus on Equating*, *2*, 115-142

Li, C., Jiao, H., & Lissitz, R. W. (2016, July). *Comparing pattern scoring with number-correct scoring in mixed-format tests.* Paper presented at the International Meeting of Psychometric Society, Asheville, NC.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30(1), 3–21.

Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions*, *19*(3), 1032.

Longabach, T. (2015). *A comparison of subscore reporting methods for a state assessment of English language proficiency.* Unpublished doctoral dissertation. University of Kansas, Lawrence, Kansas.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Reading, MA: Addison-Wesley.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29-37.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*(4), 389-404.

Maris, G., & Bechger, T. M. (2004). Equivalent MIRID models. Psychometrika, 69(4), 627-639.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Md Desa, Z. N. D. (2012). *Bi-factor multidimensional Item Response Theory modeling for subscores estimation, reliability, and classification.* Unpublished doctoral dissertation. University of Kansas, Lawrence, Kansas.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), 1-33.

Muthén, B. (1991, November). *Issues in using NAEP mathematics items to study achievement dimensionality, within-grade differences, and across-grade growth*. Paper presented at the Design and Analysis Committee of the National Assessment of Educational Progress. Washington, D. C.

Nandakumar, R. (1994). Assessing Dimensionality of a Set of Item Responses-Comparison of Different Approaches. *Journal of educational measurement*, *31*(1), 17-35.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Issues in innovative item types. In C. G. Parshall, J. Spray , J. Kalohn & T. Davey (Eds.), *Practical Considerations in Computer-Based Testing* (pp. 70–91). New York: Springer.

Plummer, M. (2017). JAGS (Version 4.3.0) [computer software]. Available from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266-285.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied psychological measurement, 9(4), 401-412.

Reckase, M. D. (1996). A linear logistic multidimensional model. In W. J. van der Linder & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 271–286). New York: Springer-Verlag.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Linden, & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer-Verlag.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D. & Hirsh, T. (1991). *Interpretation of number correct scores when the true number of dimensions assessed by a test is greater than two*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Reese, L. M. (1995). *The impact of local dependencies on some LSAT outcomes* (Report No. LSAC-R-95-02). Newtown, PA: Law School Admission Council.

Rock, D. A. (1991, November). *Subscale dimensionality*. Paper presented at the meeting of the Design and Analysis Committee of the National Assessment of Educational Progress. Washington, D.C.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349-359.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series*, *1968*(1), 1-169.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331-354.

Shin, D. (2007). *A comparison of methods of estimating subscale scores for mixed-format tests*. Pearson Educational Measurement.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28* (3), 237-247.

Skorupski, W. P. (2008, August). *A review and empirical comparison of approaches for improving the reliability of objective level scores*. Paper presented at 2008 annual meeting of A Study of the Council of Chief State School Officers. Madison, WI.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*, *70*(3), 357-375.

Smits, D. J. M., & De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research, 38*(2)*,* 161-188.

Smits, D. J. M., De Boeck, P., Verhelst, N. D., & Butter, R. (2001) The MIRID program (version 1.0) [Computer program]., Catholic University of Leuven, Belgium.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2004). *OpenBUGS user manual (Version 3.2.3)*. Cambridge, UK: MRC Biostatistics Unit.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63-86.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical software*, *12*(3), 1-16.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tanner, M. A. (1996). *Tools for statistical inference: Observed data and data augmentation* (2nd ed.). Berlin, Germany: Springer-Verlag.

Tao, J., Xu, B., Shi, N. Z., & Jiao, H. (2013). Refining the two-parameter testlet response model by introducing testlet discrimination parameters. *Japanese Psychological Research*, 55(3), 284–291.

te Marvelde, J. M., Glas, C. A., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*(1), 5-34.

Thissen, D., & Edwards, M. C. (2005). *Diagnostic scores augmented using multidimensional item response theory: Preliminary investigation of MCMC strategies*. Paper presented at annual meeting of the National Council on Educational Measurement. Montreal, Canada.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp.73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247-260.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157-186.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. V. D. Linden & C. A. W. Glas (Eds.), Computerized adaptive testing: *Theory and practice* (pp. 245–269). Dordrecht: Kluwer Academic.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores?. *Educational and Psychological Measurement*, *57*(5), 741-758.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?. *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B., Rosa, K., Nelson, L.,…Thissen, D. (2001). Augmented scores: ‘‘Borrowing strength’’ to compute scores based on small number of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220.

Wainer, H., & Wang, X. (2001). Using a new statistical model for testlets to score TOEFL. *ETS Research Report Series*, *2001*(1), i-23.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116-136.

Wang, W. C., & Jin, K. Y. (2010a). Multilevel, two-parameter, and random-weights generalizations of a model with internal restrictions on item difficulty. *Applied Psychological Measurement*, *34*(1), 46-65.

Wang, W. C., & Jin, K. Y. (2010b). A generalized model with internal restrictions on item difficulty for polytomous items. *Educational and Psychological Measurement*, *70*(2), 181-198.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126-149.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *ETS Research Report Series*, *2002*(1).

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479-494.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 267-287). New York: Springer.

Xie, C. (2014). *Cross-classified modeling of dual local item dependence* (Doctoral dissertation). Retrieved from http://hdl.handle.net/1903/15142.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83-105.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469-492.

Yen, W. M. (1984a). Obtaining Maximum Likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21*(2), 93–111.

Yen, W. M. (1984b). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance*. Paper presented at the Annual Meeting of the Psychometric Society, Montreal, Canada.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance*. Paper presented at the Annual Meeting of the Psychometric Society, Montreal, Canada.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association*, *86*(413), 79-86.

Zenisky, A. L., Hambleton, R. K., & Sired, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291-309.